User requirements and specification of the ACGT internal clinical trial

Project Number: FP6-2005-IST-026996

Deliverable id: D 2.1

Deliverable name: User requirements and specification of the ACGT internal clinical trial

Date: 13 September, 2006

COVER AND CONTROL PAGE OF DOCUMENT	
Project Acronym:	ACGT
Project Full Name:	Advancing Clinico-Genomic Clinical Trials on Cancer: Open Grid Services for improving Medical Knowledge Discovery
Document id:	D 2.1
Document name:	User requirements and specification of the ACGT internal clinical trial
Document type (PU, INT, RE)	RE
Version:	Final
Date:	13.9. 2006
Authors: Organisation: Address:	Manolis Tsiknakis FORTH - ICS Science and Technology Park of Crete, Heraklion, Greece

Document type PU = public, INT = internal, RE = restricted

ABSTRACT:

The present deliverable presents the rational of the project, in a more explicit and detailed manner, as well as its specific objectives. It elaborates the adopted requirements engineering methodology adopted by the project and shortly present details of the Clinical Studies designed for the evaluation of the results of the project.

Explicit scenarios, presenting both user-driven stories expressing user-needs, as they are documented by representative users, as well as technology-driven description of requirements of the system under design, representing indicative functionality of the system, as understood by experienced technological experts, have been developed and are described. The various future user groups and stakeholders of the project's results are presented and analyzed. Initial user requirements and functional requirements of the ACGT platform are elaborated. These requirements will be further elicited in the various WPs of the project, as foreseen in the DoW.

With the initial requirements of the project in mind, a large number of data sources, tools, standards and technology are available for the different aspects of the intended system and significant R&D results are available in a number of related scientific domains. ***PART 2*** of the deliverable provides a detailed state-of-the-art review in all of the scientific areas relevant to the project.

KEYWORD LIST: Biomedical Grid, Semantic information integration, Clinical Trials, Biomedical Ontologies, Requirements, Grid Services

MODIFICATION CONTROL			
Version	Date	Status	Author
0.1	26.6.2006	Draft	M. Tsiknakis
0.5	25.7.06	Draft	M. Tsiknakis
1.0	4.8.06	Draft	M. Tsiknakis
1.5	25.8.06	Pre-final	M. Tsiknakis
Final	10.9.06	Final	M. Tsiknakis

List of Contributors

Most of the ACGT partners have given contributions, to a lesser of greater degree, for the production of this document.

Of particular importance were the contributions of the following individuals:

- Norbert Graf, University of Saarland
- Anca Bocur, Philips Research
- Stefan Rüping, FhG-IAIS
- Thierry Sengstag - SIB
- Dimitris Kafetzopoulos, FORTH-IMBB
- Oswaldo Trelles, UMA
- Federico Garcia, UMA
- Luis Martín, UPM
- Gabriele Weiler, FhG-IBMT
- Maria Klapa, FORTH – ICEHT
- Georgios Stamatakos, ICCS
- Aran Lunzer, UHok
- Brecht Claerhout, Custodix
- Robert Belleman, UvA
- Stelios Sfakianakis, FORTH-ICS
- George Potamias, FORTH-ICS
- Nikolaus Forgo, UHANN
- Radu Gramatovici, SIVECO
- Mathias Brochhausen and Anand Kumar, IFOMIS-University of Saarland

Also many thanks are due to Norbert Graf (University of Saarland), Erwin Bonsma (Philips Research) and Andreas Persidis (BIOVISTA) who have acted as the review team and have provided important contributions towards the production of the final version of the document.

Contents

Table of Figures

Executive Summary

ACGT is an Integrated Project (IP) funded in the 6th Framework Program of the European Commission under the Action Line "*Integrated biomedical information for better health*". The high level objective of the Action Line is the development of methods and systems for improved medical knowledge discovery and understanding through integration of biomedical information (e.g. using modelling, visualization, data mining and grid technologies). Biomedical data and information to be considered include not only clinical information relating to tissues, organs or personal health-related information but also information at the level of molecules and cells, such as that acquired from genomics and proteomics research.

ACGT focuses on the domain of Cancer research, and its ultimate objective is the design, development and validation of an integrated Grid enabled technological platform in support of post-genomic, multi-centric Clinical Trials on Cancer. The driving motivation behind the project is our committed belief that the breadth and depth of information already available in the research community at large, present an enormous opportunity for improving our ability to reduce mortality from cancer, improve therapies and meet the demanding individualization of care needs.

In addition to the shear volume, data collected using a variety of laboratory technologies and techniques are often published without the background information (method of capture, sample preparation, statistical techniques applied) that is needed to reproduce results. In fact, a typical researcher spends as much time trying to understand the origins of a dataset as actually performing new analyses. Rarely is a clinical biostatistician able to make good use of data collected on studies in which they were not directly involved with, largely due to incomplete or non-existent annotation and standardization of the information. Even within a single laboratory, researchers have difficulty integrating data from different technologies because of a lack of common standards and other technological and medico-legal and ethical issues. As a result, very few cross-site studies and clinical trials are performed and in most cases it isn't possible to seamlessly integrate multi-level data. Moreover, apart from problems in sharing and re-using data, what is even more critical is the fact that clinical researchers or molecular biologists often find it hard to exploit each other's expertise due to the absence of a cooperative environment which enables the sharing of resources and tools for comparing results and experiments, and a uniform platform supporting the seamless integration and analysis of disease-related data at all levels.

The ultimate objective, therefore, of the ACGT project is the development of European Knowledge Grid infrastructure offering high-level tools and techniques for the distributed mining and extraction of knowledge from data repositories available on the Grid, leveraging semantic descriptions of components and data and offering knowledge discovery services in the domain of Cancer research. Special emphasis will be given to the trust that needs to be embedded in the platform and relevant ethical issues, thus creating optimal conditions for service uptake.

The present document is the first deliverable on the requirements of such a system. The requirements engineering process is a structured set of activities which leads to the production of a requirements document.

Due to the complexity and ample scope of the project as well as the size of the project with its many partners with complementary expertise, we primarily have to firstly define a minimal set of requirements from where to build on such a system and secondly, minimize wheel reinvention by identifying which technological components are available, and which are required to be built, in order to comply with such requirements. In the very complex domain,

within which ACGT positions itself, the requirements definition activity cannot be defined by a simple progression through, or relationship between, acquisition, expression, analysis, and specification. Requirements evolve at an uneven pace and tend to generate further requirements from the definition processes.

The project has selected indicative Clinical Trials on Cancer, namely breast cancer, pediatric nephroblastoma and in-silico modelling and simulation of tumor growth and response to treatment, for the initial requirements gathering activity. Since we see the requirements engineering process as a structured set of activities which will lead to the production of the final system requirements, an iterative requirements engineering process has been adopted, mainly based on scenarios and prototyping. Inputs to the requirements engineering process are information about existing systems, user and stakeholder needs, organizational standards, regulations and other domain information.

Explicit scenarios, presenting both user-driven stories expressing user-needs, as they are documented by representative users, as well as technology-driven description of requirements of the system under design, representing indicative functionality, as understood by experienced technological experts, have been developed. This has led to the inevitable differences in the structure and content of the presented initial scenarios, which form the basis for the acquisition of initial requirements. The project intends to continue the development of additional such scenarios, by reaching out to the wider user community that will be the future users of the ACGT technological platform, a fact already foreseen in its DoW.

Part 1 of the present deliverable presents the rational of the project, in a more explicit and detailed manner, as well as its specific objectives. It elaborates on the adopted methodology for requirements engineering and shortly present details of the Clinical Studies designed for the evaluation of the results of the project. It should be observed that the term “studies” rather than “trials” is used. The reason for this is the fact that although the presented studies are indeed formal designs of clinical trials they can not be formally implemented through the use of the ACGT technologies. The reason is that by the time the technologies are ready, there will not be adequate time to execute a properly designed Clinical Trial. Nevertheless, these properly designed clinical trials will enable the implementation of studies (i.e. reduced number of patients enrolled) which will allow the validation of the integrated ACGT technological environment as a whole but also the validation of its individual building blocks. The various future user groups and stakeholders of the project are presented and analyzed. Initial user requirements and functional requirements of the ACGT platform are elaborated. These requirements will be further elicited in the various WPs of the project, as foreseen in the DoW.

With the initial requirements of the project in mind, a large number of data sources, tools, standards and technology are available for the different aspects of the intended system and significant R&D results are available in a number of related scientific domains. ***PART 2*** of the deliverable provides a detailed state-of-the-art review in all of the scientific areas relevant to the project.

Prior to presenting the SoA in these areas, the Deliverable includes a Chapter providing a review in the domain of integrated ‘-omic’ studies for disease-specific diagnosis, prognosis and therapy design and the existing challenges. The objective is, for the readers of this document, to understand the main challenges ahead and the relevance of all of the technological domains reviewed thereafter, in truly integrated studies, i.e. studies requiring access, integration, analysis and presentation of multilevel biomedical data. Several cases and international studies are described presenting working hypothesis that the joint analysis

of genomic and proteomic data will provide more information for modelling disease susceptibility than either analysis alone.

As a result it should become clear to the reader that the ultimate objective of the foreseen ACGT technological platform is integrated access to multilevel (semantically consistent) biomedical data, its integration and analysis on the GRID.

From this point on the Deliverable provides detailed SoA review of the following scientific domains:

- **Biomedical Grids**

  Grid computing provides a novel approach to harnessing distributed resources, including applications, computing platforms or databases and file systems. Applying Grid computing can drive significant benefits by improving information access and responsiveness, and adding flexibility, all crucial components of solving the data warehouse dilemma. Special effort was made to fully review the methodological approaches taken by some key Grid projects of direct relevance to ACGT, their achievements and the technologies produced which may be of use to ACGT.

- **Systems and Standards for Clinical Trials**

  The project needs to create a number of tools for the uniform, semantically consistent reporting, and management of multi-centric clinical trials. For this reason relevant SoA is reviewed with emphasis given to approaches for “ontology based” implementation of end user applications.

- **Bioinformatics Tools**

  The number of analytical tools required to be developed, as open source services that are publishable and discoverable in the ACGT Grid, is large. On the other hand an extensive SoA review indicates that a large number of analytical tools also do exist, most of which are available as open source. This fact again identifies one of the main challenges in this domain, i.e. the utilisation of the Grid (Gridification of analytical tools) and the standardisation of the ways in which they are published on the Grid, including the metadata used to describe them, so that such tools are discovered and seamlessly orchestrated in complex analytical workflows.

  For this reason (i.e. the need for service invocation) coupled with the characteristics of the majority of the future ACGT users (i.e. non-IT experts) we have identified a need for supporting visual techniques and developing user-friendly tools supporting invocation and orchestration of services for data access, analysis, visualisation etc. A review into current SoA is, thereby, included.

  Also a Chapter has been included providing a SoA review on the use of 2D and 3D visualisation tools. In ACGT, the complexity, dimensionality and size of the data to be visualized are expected to increase rapidly. This calls for a flexible and powerful visualization environment. Apart from the needs for the 3D visualization of the results of the Oncosimulator there are needs for such tools to be used for the interactive 3D visualization of scatter plots, Principal Component Analyses, 3D graphs/networks and perhaps even 3D self-organizing maps.

- **Ontologies and Semantic Information Integration**

A central component of the foreseen ACGT platform is its Master Ontology. The ACGT Master Ontology on Cancer has a pivotal role for enabling (a) semantic information integration and (b) ontology based implementation of end-user applications. As a result SoA reviews are included both in the domain of Biomedical Ontologies as well as Semantic Integration of Heterogeneous Databases.

- ***Metadata***

ACGT focuses on the semantic integration of data but also on the discovery, integration, and management of sharable data assets (i.e. data and tools operating on such data). As a result the issue of metadata becomes of paramount importance for the successful achievement of the project objectives. Metadata is often called 'data about data'. More precisely, it is the underlying definition or structured description of the content, quality, condition or other characteristics of data. The relevant Chapter provides a SoA review, presents related standards and experiences of other international projects with somewhat similar to ACGT requirements and objectives.

- ***Workflows***

In order to support scientists in their data management and analysis tasks, scientific workflows have recently gained increased interest and momentum as a unifying mechanism for handling scientific data. Scientific workflows pose a unique set of challenges, due to the nature of scientific data and the specific needs for large-scale data collection, querying and analysis.

The ultimate objective of the project is the delivery of a user-friendly workflow environment to its users, enabling efficient discovery of data sources and analytical tools as well as their seamless integration into scientific workflows and their execution on the Grid. Relevant SoA is presented, available workflow platforms are evaluated and the experiences of relevant to ACGT projects are analyzed, especially with respect to the need for metadata description of the workflows themselves.

- ***In-silico modelling and Simulation***

Also, a detailed SoA in the domain of In-silico modelling and simulation of tumour response to therapy is included, to be used as the foundation for the selection and specification of the required technologies and scientific methods for the ACGT In-Silico modelling and simulation trial.

The present document also includes detailed SoA of the current Legal and Ethical issues involved in multi-centric post-genomic Clinical Trials, the current regulatory framework and of techniques for privacy enhancing. These domains will enable the specification of the ***security view*** of the ACGT architecture and the implementation of the, domain specific, security services of the platform.

Finally, Grid portal technologies and on-line training platforms are reviewed, since the ACGT Grid portal will provide the point of access to the ACGT infrastructure and also be used for the continues training of the wider community addressed by ACGT.

PART 3 of the Deliverable includes a number of Appendices, with additional very useful information on various topics and an elaborate list of abbreviations and acronyms.

PART 1

User Needs and Requirements

1 Introduction

1.1 Project background

Recent and forthcoming developments in genomics and the increased importance of genetics in healthcare are already changing clinical care. Research on the molecular mechanisms of cell growth, apoptosis and differentiation has resulted in a better understanding of the nature of cancer cells. The genotypic knowledge of a cancer cell helps to identify the predisposition of the disease and develops therapies adapted to the genotype of a cancer patient. Medicine is getting more individualised.

The information from genetic and protein studies, clinical trials, and other research is growing rapidly. On the other hand there is no unifying infrastructure or common standard for the technologies that cancer researchers use. There are for example no mechanisms for easily sharing and joining data. In responding to these challenges, Biomedical Informatics is quickly evolving into a research field that encompasses the use of all kinds of biomedical information, from genetic and proteomic data to image and clinical data associated with various levels of the human body. This kind of integration and exploitation of the data and information requires a new synergetic approach that enables a bi-directional dialogue between these scientific disciplines and integration in terms of data, methods, technologies, tools and applications. While the goal is clear, the path is difficult to go, fraught with technical, scientific, clinical, legal and ethical challenges. Many new tools for today's biomedical researcher have been developed to find the mechanism behind cancer, whereas legal and ethical issues are lagging behind.

The main objective of ACGT is the fight against cancer. To achieve this goal ACGT has the following objectives

- The ACGT project sees its mission to develop a GRID platform to support and stimulate further exchanges of both clinic and genetic information.
- ACGT intends to trigger the emergence of latent ***clinico-genomic synergies*** to ensure ***faster diagnosis*** and more ***efficient therapy***
- ACGT targets two major cancer diseases namely, ***breast cancer*** *(BRCA)* and ***paediatric nephroblastoma*** *(PN)* presented by three (running) clinical trials.
- In addition, ***in-silico*** oncology trial scenarios will be run to assess the utility of ***tumour-growth simulation*** on both BRCA and PN.

1.2 The ACGT Environment

ACGT was set up to respond to the challenges arising from three global factors as mentioned above:

- Changing environment comprising a number of issues in all areas of life science
- Changes in healthcare delivery comprising the move towards a more individualised medicine and
- Technology push in conjunction with Biomedical Informatics

ACGT focuses on clinical trials on Cancer (Wilms tumor, Breast) and is based on the principles of open access (among trusted partners) developing open source products.

ACGT will provide a unified technological infrastructure to facilitate

- integrated access to multi-level biomedical data;
- The development or re-use of open source analytical tools, accompanied with the appropriate meta-data allowing their discovery and orchestration into complex workflows.

ACGT brings together internationally recognised leaders in their respective fields, with the aim to deliver to the cancer research community an integrated clinicogenomic ICT environment enabled by a powerful GRID infrastructure. In achieving this objective ACGT has formulated a coherent, integrated workplan for the design, development, integration and validation of all technologically challenging areas of work. Namely:

- Grid: delivery of a European Biomedical Grid infrastructure offering seamless mediation services for sharing data and data-processing methods and tools, and advanced security;
- Integration: semantic, ontology based integration of clinical and genomic/proteomic data - taking into account standard clinical and genomic ontologies and metadata;
- Knowledge Discovery: delivery of data-mining GRID services in order to support and improve complex knowledge discovery processes;
- Legal and ethical issues: development and integration of technical solutions regarding data protection and secure personal data management in a European context.

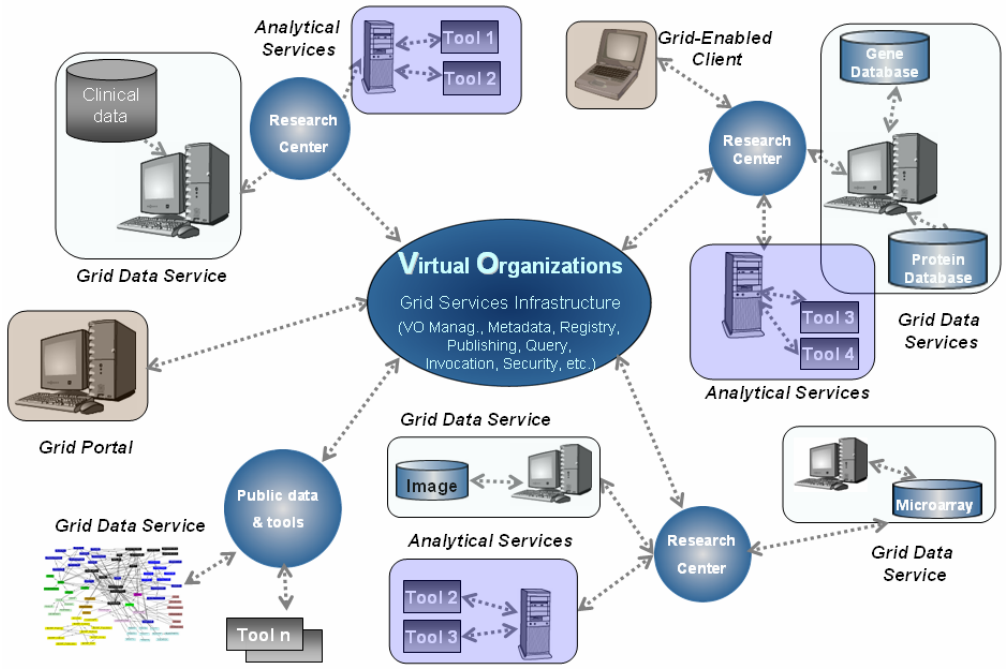


Figure 1: The ACGT Virtual Organizations

The technological platform of ACGT will be validated in concrete setting of advanced clinical trials on Cancer. Pilot trials have been selected based on the presence of clear research objectives, raising the need to integrate data at all levels of the human being.

1.3 Vision of the Project

Information arising from post-genomics research, and combined genetic and clinical trials on one hand, and advances from high-performance computing and informatics on the other is rapidly providing the medical and scientific community with new insights, answers and capabilities. The breadth and depth of information already available in the research community at large, present an enormous opportunity for improving our ability to reduce mortality from cancer, improve therapies and meet the demanding individualization of care needs.

⇨ *A critical set of challenges, however, currently inhibit our capacity to harvest these opportunities.*

Capitalizing on the opportunities apparent to cancer research community requires the integration and exploitation of data and information generated at all levels (from molecular to organ and disease to the population) by the disciplines of bioinformatics and medical informatics, including medical imaging. So, the new raising biomedical informatics (BMI) technology will be utilised and accordingly enhanced in order to meet the posted needs and technological challenges.

This, in turn, requires a new synergetic approach that enables a bi-directional collaboration among these scientific disciplines and integration in terms of data, methods, technologies, tools and applications. In the new area of "genomic medicine", designers and analysts of clinical trials must consider various issues, such as:

- how to design experiments for obtaining coherent and consistent medical and biological data, while avoiding various types of biases and errors,
- how to develop methods for heterogeneous (e.g., genomic, medical) data source integration, including the use of ontologies which facilitate mapping and information retrieval,
- how to develop methods for data selection, checking, cleaning, and pre-processing of combined genomic/medical data, and
- How to incorporate collaborative approaches to data analysis, since biomedical statisticians and data miners in genomics and medicine have been following different methodologies, and dedicated, often proprietary, tools.

In order to address all the previously stated issues, there is an obvious need for an integrated and high-performing Grid-enabled Clinico-Genomic Environment that will not only allow the ubiquitous access to computational resources but it will also enhance the cross-organizational resource sharing. The ultimate objective therefore of the ACGT project is the development of European Knowledge Grid infrastructure offering high-level tools and techniques for the distributed mining and extraction of knowledge from data repositories available on the Grid, leveraging semantic descriptions of components and data and offering knowledge discovery services in the domain of biomedical informatics. Special emphasis will be given to the trust that needs to be embedded in the platform and relevant ethical issues, thus creating optimal conditions for service uptake. Most importantly, such an environment will facilitate the design, monitoring, and evaluation of clinical trials and experimental treatment protocols.

1.4 The ACGT specific objectives

In order to achieve its vision and goals, the project will create and test an infrastructure for cancer research by using a:

> *virtual web of trusted and interconnected organizations and individuals to leverage the combined strengths of cancer centers and investigators and enable the sharing of biomedical cancer-related data and research tools in a way that the common needs of interdisciplinary research are met and tackled.*

A major part of the project is devoted to research and development in infrastructure components that eventually will be integrated into a workable demonstration platform upon which the selected use cases (the Clinical Pilots) can be demonstrated and evaluated against user requirements defined at the onset of the project.

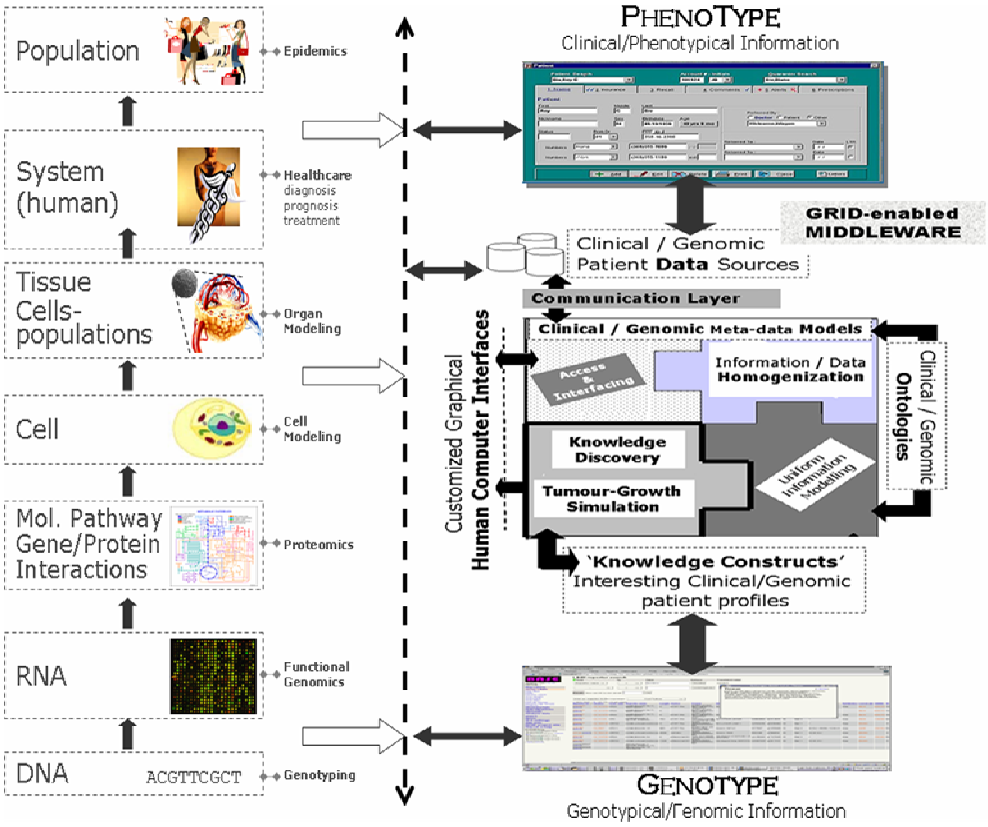


Figure 2: The envisioned ACGT Grid-enabled infrastructure and integrated environment – integration to be achieved at all levels, from the molecular to system and to the population.

In Figure 2 (previous page), an outline schema of the foreseen *ACGT* infrastructure and integrated clinico-genomic environment is shown. Integration targets all levels – from molecular to the human and the population. Grid-enabled mediation functionality (realised by respective services) enable knowledge-enriched, effective and reliable integration.

The principal ***user service objectives*** of the ACGT project can be summarized as:

- ⇨ Secure that the services reflects the identified stakeholder priorities and validate them in reference to user priorities, i.e. functionality, security and user acceptance.

- ⇨ Validate the economic feasibility of sustainability of the platform and elaboration of alternative exploitation models, bearing in mind the open source-open access principles on which the project was build.
- ⇨ Assess the potential usage of the ACGT platform in other domains.

The principal ***business objectives*** of the ACGT project can be summarized as:

- ⇨ Develop suitable business models based on value creation and value nets.
- ⇨ Contribute to the creation of a "culture" of sharing of biomedical data and tools based upon common standards and their utilisation in a manner that protects data privacy and security.
- ⇨ Build a critical mass for the commercial exploitation of ACGT.

The principal ***technological objectives*** of the ACGT project can be summarized as:

- ⇨ Integration of Clinical Research Centers on Cancer with varying needs and capabilities in a common network for sharing data, applications, and technologies.
- ⇨ Development of a useable and scalable biomedical Grid that Clinical Research Centers on Cancer will actively use for added value clinical trials.
- ⇨ Development of new open source tools for multilevel, biomedical data analysis and knowledge discovery as well as adaptation/modification of existing ones so that they comply with the ACGT integration guidelines and architecture, utilise the advantages of Grid computing, and enable high-performing data-mining and biomedical knowledge extraction operations.
- ⇨ Development of a Master Ontology for Clinical Trials and Cancer
- ⇨ Enable ontology-based implementation of key Clinical Trial Management Modules.
- ⇨ Utilisation of clinical trial management systems based on standards-based and components-based clinical trial management systems, integrative cancer research applications and innovative tools to support (a) ontology-based integration and sharing of data and biomedical information and (b) advanced data mining and biomedical knowledge extraction.
- ⇨ Sharing of biomedical information and data upon common standards and utilisation of and in a manner that protects data privacy and security.
- ⇨ Fostering common usage of vocabularies, common data elements and the formation of a unifying architecture for the support of the advanced clinico-genomics clinical trials of the future.
- ⇨ Validate the technical performance of the platform to ensure that it fully supports the identified business needs at acceptable performance.

In achieving these objectives the main technological challenges of the project are:

- **Semantic Grid** services, enabling large-scale (semantic, structural, and syntactic) interoperation among biomedical resources and services;
- Master **ontology** (on Cancer) through semantic modelling of biomedical concepts using existing ontologies and ontologies developed for the needs of the project;
- Open source bioinformatics tools and other **analytical services**;
- Semantic **annotation** and **advertisement** of biomedical resources, to allow **metadata-based discovery** and query of tools, and services;

- Orchestration of data access and analytical services into **complex eScience workflows** for post genomic clinical research and trials on cancer;

- **Meta-data descriptions of clinical trials** to provide adequate provenance information for future re-use, comparison, and integration of results;

1.5 Purpose and Structure of this Document

The present deliverable is the D2.1: User Requirements and specifications of the ACGT internal clinical trial". State-of-the-art analysis for the ACGT project is the result of part of the work performed in work tasks T2.1- State-of-the-art Review, T2.2- User Needs and Requirements Analysis and T2.3 – Scenario based Requirements for ACGT trials. It represents a major project Milestone (M2).

The purpose of the work done and reported here is to:

- ✧ develop user scenarios as a methodological approach to gathering and eliciting user requirements and

- ✧ perform a state-of-the-art analysis in all scientific domains relevant to the project.

The results reported in the present deliverable will be input to the following work on functional, security and trust and societal user requirements specifications.

The purpose of the present state-of-the-art-analysis is to map out existing and future technologies and methodologies, which will be available for the prototype platform and used in the validation. Even though future technologies may create opportunities for new services on the ACGT platform, emphasis has been put on identifying existing technologies, which can deliver the desired functionality.

The present deliverable is articulated as follows:

Part 1 of the deliverable presents the rational of the project, in a more explicit and detailed manner, as well as its specific objectives. It elaborates on the adopted requirements engineering methodology and shortly presents details of the Clinical Studies designed for the evaluation of the results of the project. It should be observed that the term "studies" rather than "trials" is used. The reason for this is the fact that although the presented studies are indeed formal designs of clinical trials they can not be formally implemented through the use of the ACGT technologies. The reason is that by the time the technologies are ready, there will not be adequate time to execute a properly designed Clinical Trial. Nevertheless, these properly designed clinical trials will enable the implementation of studies (i.e. reduced number of patients enrolled) which will allow the validation of the integrated ACGT technological environment as a whole as well as validation of its individual building blocks.

With the initial requirements of the project in mind, a large number of data sources, tools, standards and technology are available for the different aspects of the intended system. ***PART 2*** of the deliverable provides a detailed state-of-the-art review in all of the scientific areas relevant to the project.

2 Adopted methodology for the Engineering of User Requirements

2.1 Introduction

System engineering is one of the most critical aspects of successful, large-scale, software-intensive, mission-critical system development, and post-deployment support. It is a process that commences with the earliest phases of the system lifecycle and continues until the system is retired from use. System engineering involves a complex series of activities necessary to:

- Transform an operational need into a description of system performance parameters and a preferred system configuration using an iterative process of functional analysis, synthesis, optimization, definition, design, test, and evaluation;
- Integrate related technical parameters and assure compatibility of all physical, functional, and program interfaces in a manner that optimizes the total system definition and design; and
- Integrate performance, reliability, maintainability, manageability, supportability, and other factors into the total engineering effort.

Regardless of the application domain, engineering large software-intensive systems is a complex activity that typically involves hundreds, if not thousands, of geographically distributed engineers representing numerous specialties, e.g., hardware, software, human factors, mission analysis, sensors, supportability, reliability and maintainability.

The current Chapter of this deliverable aims to explain the process and the methodology adopted for gathering and eliciting user and system requirements.

2.2 System engineering activities

The system engineer has the critical role--responsibility for transforming user or mission requirements into a real, cost-effective system. The various activities that compose the System requirements process according to the IEEE 1471, 2000: IEEE Recommended Practice for Architectural Description of Software-Intensive Systems is:

- Requirements Engineering.
- System Design and Allocation.
- Interface Definition and Integration.
- Engineering Decision-Making.
- Change Impact Analysis and Management.
- Integration Planning and Management.
- Quality Engineering and Assurance.

The various Requirements Engineering activities are:

- **User requirements** - The requirements for the system as a whole are established and written to be understandable to all stakeholders.
- **Architectural design** - The system is decomposed into sub-systems.
- **Requirements partitioning** - Requirements are allocated to these sub-systems. A critical task in any component-based system or a Service Oriented System Architecture (also see: International workshop on Service Oriented System Engineering 2005 - http://asusrl.eas.asu.edu/srlab/activities/sose2005/SOSA)
- **Software requirements engineering** - More detailed system requirements are derived for the system software.
- **Sub-system development** - The hardware and software sub-systems are designed and implemented in parallel.
- **System integration** - The hardware and software sub-systems are put together to make up the system.
- **System validation** - The system is validated against its requirements.

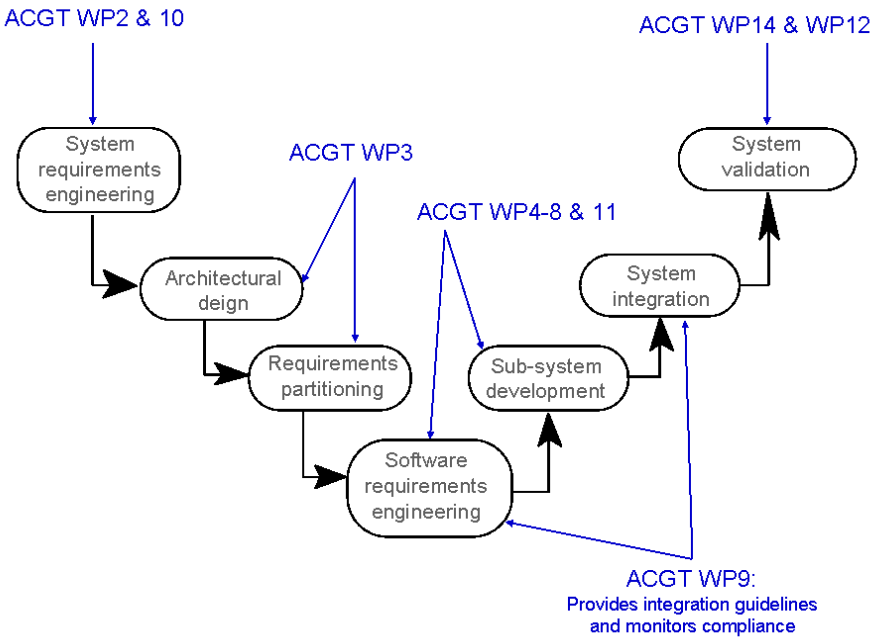


Figure 3: The decomposition of the various RE tasks in the WPs of the ACGT Workplan

These Requirements Engineering activities and steps are allocated to separate Work Packages in the ACGT workplan (see Fig. 3), with WP2 - which produces the current deliverable - been responsible to select, elaborate and document the “initial” requirements for the system (form the user’s point of view).

2.3 System requirements engineering

The requirements engineering process is a structured set of activities which lead to the production of a requirements document. Inputs to the requirements engineering process are information about existing systems, stakeholder needs, organizational standards, regulations and domain information. Requirements engineering processes include requirements elicitation, requirements analysis and negotiation and requirements validation.

Requirements engineering process models are simplified process descriptions which are presented from a particular perspective. Human, social and organizational factors are important influences on requirements engineering processes (see: IEEE/ANSI 830-1993 standard).

The goal of requirements engineering is the production of a good requirements specification. The IEEE Guide to Software Requirements Specifications defines a good software requirements specification as being:

- unambiguous
- complete
- verifiable
- consistent
- modifiable
- traceable
- usable during operations and maintenance

Recent requirements engineering literature is in agreement on this set of attributes, with the added property that the requirements should be necessary. The requirements should be prioritized as well, particularly in novel situations where the order in which the sub goals are addressed significantly impacts the final solution [POT2001].

Conversely, problems with requirements elicitation inhibit the definition of requirements which are unambiguous, complete, verifiable, consistent, modifiable, traceable, usable, and necessary. Some of these problems are looked at in the next section. In subsequent sections the process sketched out here as “fact-finding, information gathering, and integration” will be refined to specifically address the problems encountered during requirements elicitation.

2.4 Requirements Elicitation

Rzepka decomposes the requirements engineering process into three activities [RZE1989]:

1. elicit requirements from various individual sources;
2. insure that the needs of all users are consistent and feasible; and
3. validate that the requirements so derived are an accurate reflection of user needs.

This model implies a sequential ordering to the activities, with elicitation done once at the very beginning of the process. In reality, though, the process is iterative (see figure 4), with these activities revisited many times:

> ...the requirements definition activity cannot be defined by a simple progression through, or relationship between, acquisition, expression, analysis, and

specification. Requirements evolve at an uneven pace and tend to generate further requirements from the definition processes.

The construction of the requirements specification is inevitably an iterative process which is not, in general, self-terminating. Thus, at each iteration it is necessary to consider whether the current version of the requirements specification adequately defines the user's requirement, and, if not, how it must be changed or expanded further.

Thus, while requirements elicitation consists of the earliest activities in the requirements engineering process, it can not be divorced from the subsequent activities. Elicitation will likely iterate with these other activities during requirements development.

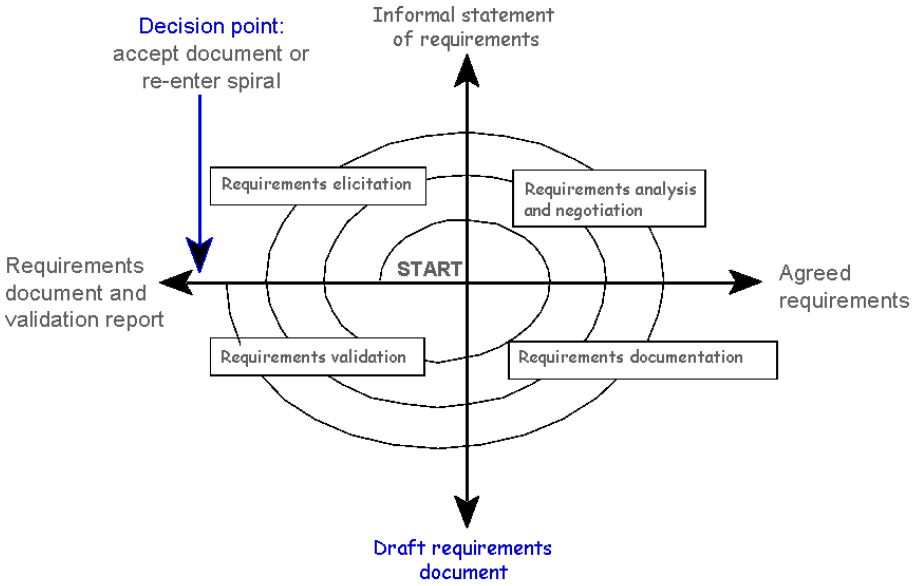


Figure 4: Spiral Model of the RE process

Requirements elicitation itself can be broken down into the activities of fact-finding, information gathering, and integration. For example, Rzepka further decomposes the elicitation process as follows [RZE1989]:

1. Identify the relevant parties which are sources of requirements. The party might be an end user, an interfacing system, or environmental factors.

2. Gather the "wish list" for each relevant party. This wish list is likely to originally contain ambiguities, inconsistencies, non-feasible and non-testable requirements. This original "wish list" may also prove to be incomplete.

3. Document and refine the "wish list" for each relevant party. The wish list includes all important activities and data, and during this stage it is repeatedly analyzed until it is self-consistent. The list is typically high level, specific to the relevant problem domain, and stated in user-specific terms.

4. Integrate the wish lists across the various relevant parties, henceforth called viewpoints, thereby resolving the conflicts between the viewpoints. Consistency checking is an important part of this process. The wish lists, or goals, are also checked for feasibility.

5. Determine the non-functional requirements, such as performance and reliability issues, and state these in the requirements document.

These activities are common to most of the process definitions for requirements elicitation found in the literature. However, the means of achieving these activities and iterating between them are still not well understood.

The resulting product from the elicitation phase is a subset of the goals from the various parties which describe a number of possible solutions. The remainder of the requirements engineering process concerns the validation of this subset to see if it is what the sponsor/funder and user actually intended. This validation typically includes the creation of models to foster understanding between the parties involved in requirements development. The result of a successful requirements engineering process is a requirements specification, where “the goodness or badness of a specification can be judged only relative to the user’s goals and the resources available”.

In general the core Requirements Engineering activities can be grouped into the following main categories:

- eliciting requirements,
- modelling and analyzing requirements,
- communicating requirements,
- agreeing requirements, and
- evolving requirements.

2.4.1 Specific elicitation techniques

A number of techniques exist and are often used during the requirements elicitation phase of a project. They include:

- Interviews
- Scenarios
- Soft systems methods
- Prototyping
- Observations and social analysis
- Requirements reuse

The complexity of the domain which is addressed by the ACGT project necessitated that a spiral process of requirements analysis, elicitation, documentation and validation is adopted. Specific techniques have also been selected for the elicitation, negotiation and agreement of requirements as well as their validation. These techniques are ***scenarios*** and ***prototyping***.

2.4.1.1 Scenarios

"Scenarios are arguably the starting point for all modelling and design" [THO1992]. They allow us to take a backward glance. They use a simple, traditional activity – storytelling – to

provide a vital missing element, namely a view of the whole of a situation. 'A straight-line sequence of steps taken by independent agents playing system roles' is roughly what most engineers mean by Scenario. Synonyms include Operational Scenario – itself part of a Concept of Operations, (Test) Case (of actions or events). A sequence of numbered steps in the form '<role> does <action>' is a simple and effective way to describe how a result is to be obtained.

Scenarios are a powerful antidote to the complexity of system development [ROL1998]. Telling stories about systems helps to ensure that project stakeholders share a sufficiently wide view to avoid missing vital aspects of problems. Scenarios vary from brief stories to richly-structured analyses, but are almost always based on the idea of a sequence of actions carried out by intelligent agents. People are very good at reasoning from even quite terse stories, detecting inconsistencies, omissions, and threats with little effort. These innate human capabilities give scenarios their power. Scenarios are applicable to systems of all types, and may be used at any stage of the development life-cycle for different purposes [SUT2003].

Scenarios are stories which explain how a system might be used. They should usually include:

- a description of the system state before entering the scenario;
- the normal flow of events in the scenario;
- exceptions to the normal flow of events;
- information about concurrent activities;
- a description of the system state at the end of the scenario.

2.4.1.2 Prototypes

A prototype is an initial version of a system which may be used for experimentation. Prototypes are valuable for requirements elicitation because users can experiment with the system and point out its strengths and weaknesses. The prototype allows users to experiment and discover what they really need to support their work, but more importantly forces a detailed study of the requirements which reveals inconsistencies and omissions.

Rapid development of prototypes is essential so that they are available early in the elicitation process. Establishes feasibility and usefulness before high development costs are incurred.

Prototypes are also essential for developing the ‘look and feel’ of a user interface and they can be used for system testing and the development of documentation.

A prototype of a proposed system is presented to workers for critical comments. Revisions are made to the original prototype, producing a second version that is again presented to users for critical analysis. The process of revising and submitting to users continues until some criterion for acceptability is reached. [THO1992] advocates the use of rapid prototyping as a vehicle for allowing technical communicators to become a part of the development team.

2.4.1.3 Volatility of Requirements

Requirements change. During the time it takes to develop a system the users’ needs may mature because of increased knowledge brought on by the development activities, or they

may shift to a new set of needs because of unforeseen organizational or environmental pressures.

If such changes are not accommodated, the original requirements set will become incomplete, inconsistent with the new situation, and potentially unusable because they capture information that has since become obsolete. One primary cause of requirements volatility is that "user needs evolve over time" [SAG1990]. The requirements engineering process of elicit, specify, and validate should not be executed only once during system development, but rather should be returned to so that the requirements can reflect the new knowledge gained during specification, validation, and subsequent activities. A requirements engineering methodology should be iterative in nature, "so that solutions can be reworked in the light of increased knowledge" [SUD2005].

Another cause of requirements volatility is that the requirements are the product of the contributions of many individuals, and these individuals often have conflicting needs and goals. For example, there usually is more than one stakeholder, with each one having different and often contradictory views and interests. Due to political climate and other factors, the needs of a particular group may be overemphasized in the elicitation of requirements. Later prioritization of the elicitation communities' needs may correct this oversight and result in requirements changes. Both the traceability of requirements and their consistency may be affected if these changes are frequent and not anticipated.

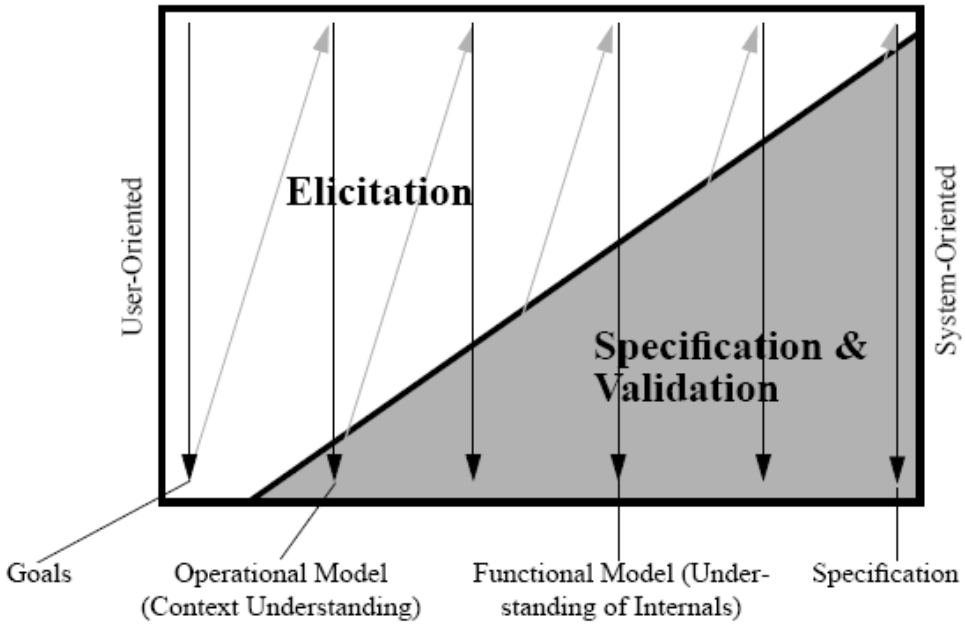


Figure 5: Requirements Engineering as an Iterative Process

Organizational complexity is another cause of requirements volatility. Organizational goals, policies, structures, and work roles of intended end users all may change during the course of a system's development, especially as the number of users affected by a system's development increases. An iterative process of requirements development can address the problems of volatility [SUT2002]:

> The traditional notion of the software development life-cycle with requirements capture being completed before the design stage is no longer satisfactory. Requirements capture and design are now seen to be symbiotic. The initial set of requirements needed to start off the design process is

gradually refined into a systematic and coherent statement of requirements hand in hand with the refinement of design.

Due to the problems of understanding and scope, user needs may not be clearly expressed initially in the requirements, and the developer or requirements analyst may make some incorrect assumptions based on this ambiguity. With an iterative process (see Fig. 5), those mistaken assumptions can be detected faster and corrected sooner.

2.5 References

[ROL1998] Rolland C., Souveyet C., Ben Achour C., *Guiding Goal Modelling Using Scenarios*. IEEE Transactions on Software Engineering, special issue on Scenario Management, 1998.

[POT2001] Potts C. Metaphors of intent. In Proceedings of the fifth IEEE International Symposium on Requirements Engineering,2001, 31-38

[SUT1998] Sutcliffe A, et al., Supporting Scenario-Based Requirements Engineering, IEEE Transactions on Software Engineering, vol. 24, no. 12, pp. 1072-1088, December, 1998

[SUT2002] Sutcliffe A., User-Centred Requirements Engineering - Theory and Practice, Springer, 2002, ISBN: 1852335173

[SUT2003] Sutcliffe A., Scenario-Based Requirements Engineering, p. 320, 11th IEEE International Requirements Engineering Conference (RE'03), 2003.

[THO1992] Thompson, M. & Wishbow, N. "Improving Software and Documentation Quality Through Rapid Prototyping." In Proceedings of the 10th Annual International Conference on Systems Documentation (1992).

[SUD2005] Sudhakar M., Managing the Impact of Requirements Volatility, Master Thesis, Department of Computing Science, Umeå University, Sweden, 2005, http://www.cs.umu.se/education/examina/Rapporter/MundlamuriSudhakar.pdf

[FIC1988] Fickas, Stephen, and Nagarajan, P. Critiquing Software Specifications, IEEE Software 5:37-47, November 1988.

[RZE1989] Rzepka, William E. A Requirements Engineering Testbed: Concept, Status, and First Results. In Bruce D. Shriver (editor), Proceedings of the Twenty-Second Annual Hawaii International Conference on System Sciences, 339-347. IEEE Computer Society, 1989.

[SAG1990] Sage, Andrew P., and Palmer, James D. Software Systems Engineering. New York: John Wiley & Sons, 1990.

3 Clinical Trials

3.1 What is clinical research?

Clinical research enables doctors and researchers to find new and better ways to understand, detect, control and treat illness. A clinical research study is a way to find answers to difficult scientific or health questions.

The most commonly performed clinical trials evaluate new drugs, medical devices, biologics, or other interventions to patients in strictly scientifically controlled settings, and are required for regulatory authority approval of new therapies. Trials may be designed to assess the safety and efficacy of an experimental therapy, to assess whether the new intervention is better than standard therapy, or to compare the efficacy of two standard or marketed interventions. The trial objectives and design are usually documented in a Clinical trial protocol.

A protocol is a study plan on which all clinical trials are based. The plan is carefully designed to safeguard the health of the participants as well as answer specific research questions. A protocol describes what types of people may participate in the trial; the schedule of tests, procedures, medications, and dosages; and the length of the study. While in a clinical trial, participants following a protocol are seen regularly by the research staff to monitor their health and to determine the safety and effectiveness of their treatment.

3.1.1 Types of Clinical Trials

The study design that provides the most compelling evidence of a causal relationship between the treatment and the effect is the ***randomized controlled trial***. Studies in epidemiology such as the ***cohort study*** and the ***case-control study*** are clinical studies in that they involve human participants, but provide less compelling evidence than the randomized controlled trial. The major difference between clinical trials and epidemiological studies is that, in clinical trials, the investigators manipulate the administration of a new intervention and measure the effect of that manipulation, whereas epidemiological studies only observe associations (correlations) between the treatments experienced by participants and their health status or diseases. These are fundamental distinctions in evidence-based medicine.

Currently some clinical trials are designed to be randomized, double-blind, and placebo-controlled. This means that each study subject is randomly assigned to receive one of the treatments, which might be the placebo. Neither the subjects nor scientists involved in the study know which study treatment is being administered to any given subject; and, in particular, none of those involved in the study know which subjects are being administered a placebo.

While the term clinical trial is most commonly associated with large randomized studies, many clinical trials are relatively small. They may be "sponsored" by single physicians or a small group of physicians, and are designed to test simple questions. Other clinical trials require large numbers of participants followed over long periods of time, and the trial sponsor is more likely to be a commercial company or a government, or other academic, research body. It is sometimes necessary to organize multicentric trials. Often the centres taking part

in such trials are in different countries (in which case they may be termed international clinical trials).

The number of patients enrolled in the study also has a large bearing on the ability of the trial to reliably detect an effect of a treatment. This is described as the "power" of the trial. It is usually expressed as the probability that, if the treatments differ in their effect on the outcome of interest, the statistical analysis of the trial data will detect that difference. The larger the sample size or number of participants, the greater the statistical power. However, in designing a clinical trial, this consideration must be balanced with the greater costs associated with larger studies. The power of a trial is not a single, unique value; it estimates the ability of a trial to detect a difference of a particular size (or larger) between the treated and control groups.

Clinical trials can also be categorized according to the type of sponsor for the trial. Investigator initiated trials have to be distinguished from trials that are sponsored by pharmaceutical companies. Especially the investigator initiated trails are lacking logistic and financial support today. They are not commercially funded. Most of these trials are clinical trials for the optimization of treatments in a clinical setting.

Another differentiation of trials can be done by their primary task. This dividese them into the following types:

Treatment trials test experimental treatments, new combinations of drugs, or new approaches to surgery or radiation therapy.

Prevention trials look for better ways to prevent disease in people who have never had the disease or to prevent a disease from returning. These approaches may include medicines, vitamins, vaccines, minerals, or lifestyle changes.

Diagnostic trials are conducted to find better tests or procedures for diagnosing a particular disease or condition.

Screening trials test the best way to detect certain diseases or health conditions.

Quality of Life trials (or Supportive Care trials) explore ways to improve comfort and the quality of life for individuals with a chronic illness

3.1.2 Phases of clinical trials

Pharmaceutical clinical trials are commonly classified into four phases. The drug-development process will normally proceed through all four stages over many years. If the drug successfully passes through the first three phases, it will usually be successfully approved for use in the general population.

Before pharmaceutical companies start clinical trials on drugs, extensive pre-clinical studies are conducted.

1. ***Phase I***

   Phase I trials are the first-stage of testing in human subjects. Normally a small (20-80) group of healthy volunteers will be selected. This phase includes trials designed to assess the safety, (Pharmacovigilance), tolerability, pharmacokinetics, and pharmacodynamics of a therapy. These trials are almost always conducted in an inpatient clinic, where the subject can be observed by full-time medical staff. The subject

is usually observed until several half-lives of the drug have passed. Phase I trials also normally include dose-ranging studies so that doses for clinical use can be refined. The tested range of doses will usually be a small fraction of the dose that causes harm in animal testing. Phase I trials most often include healthy volunteers, however there are some circumstances when patients are used, such as with oncology and HIV drug trials. In Phase I trials of new cancer drugs, for example, patients with advanced (metastatic) cancer are used. These trials are usually offered to patients who have had other types of therapy and who have few, if any, other treatment choices.

There are two specific kinds of Phase I trials – Single Ascending Dose (SAD) studies, and Maximum Tolerated Dose (MAD) studies.

SAD - Single Ascending Dose studies are those in which groups of three or six patients are given a small dose of the drug and observed for a specific period of time. If they do not exhibit any adverse side effects, a new group of patients is then given a higher dose. This is continued until intolerable side effects start showing up, at which point the drug is said to have reached the Maximum Tolerated Dose (MTD).

MAD - Multiple Ascending Dose studies are conducted to better understand the pharmacokinetics/pharmacodynamics of the drug. In these studies, a group of patients receives a low dose of the drug and the dose is subsequently escalated up to a predetermined level. Samples of blood and other fluids are collected at various time points and analyzed to understand how the drug is processed within the body.

2. Phase II

Once the initial safety of the therapy has been confirmed in Phase I trials, Phase II trials are performed on larger groups (100-300) and are designed to assess clinical efficacy of the therapy; as well as to continue Phase I assessments in a larger group of volunteers and patients. The development process for a new drug commonly fails during Phase II trials due to the discovery of poor efficacy or toxic effects.

Phase II studies are sometimes divided into Phase IIA and Phase IIB. Phase IIA is specifically designed to assess dosing requirements, whereas Phase IIB is specifically designed to study efficacy.

Some trials combine Phase I and Phase II into a single trial, monitoring both efficacy and toxicity.

3. Phase III

Phase III studies are large double-blind randomized controlled trials on large patient groups (1000-3000 or more) and are aimed at being the definitive assessment of the efficacy of the new therapy, especially in comparison with currently available alternatives. Phase III trials are the most expensive, time-consuming and difficult trials to design and run; especially in therapies for chronic conditions. Once a drug has proven satisfactory over Phase III trials, the trial results are usually combined into a large document containing a comprehensive description of the methods and results of human and animal studies, manufacturing procedures, formulation details, and shelf life. This collection of information makes up the "regulatory submission" that is provided for review to various regulatory authorities in different countries, such as the European Medicines Agency (EMEA) for marketing approval.

It is also common practice with many drugs whose approval is pending, that certain phase III trials will continue in an attempt at "label expansion." In other words, proving additional efficacy for uses beyond the original use for which the drug was designed. While not required in all studies, it is typically expected that there be at least two successful phase III trials, proving a drug's safety and efficacy, for approval from the standard regulatory agencies.

4. Phase IV

Phase IV trials involve the post-launch safety surveillance and ongoing technical support of a drug. Phase IV studies may be mandated by regulatory authorities or may be undertaken by the sponsoring company for competitive or other reasons. Post-launch safety surveillance is designed to detect any rare or long-term adverse effects over a much larger patient population and timescale than was possible during the initial clinical trials. Such adverse effects detected by Phase IV trials may result in the withdrawal or restriction of a drug.

3.2 International Guidelines in Clinical Research Studies and Trials

The value of clinical trials as the optimum methodology for the testing and evaluation of new treatments and medicines is well recognised within the research community. Clinical trials conducted on human participants are designed and conducted according to sound scientific and ethical standards within the framework of good clinical practice. Compliance with these standards provides the public with assurance that the rights, safety and well being of trial participants are protected and that the clinical trial data are credible. Guidelines for clinical trials should be read in the context of the Declaration of Helsinki, October 2000 and the ICH Harmonised Tripartite Guidelines for Good Clinical Practice, May 1997.

A European Union (EU) Directive was published in May 2001 – The EU Directive on Clinical Trials (2001/20/EC) [DIR2001]. The EU clinical trials directive regulates the conduct of clinical trials involving medicines for human use. The aims of the directive are:

- To protect the rights, safety and well being of trial participants
- To simplify and harmonise the administrative provisions governing clinical trials
- To establish a transparent procedure that will harmonise trial conduct in the EU and ensure the credibility of results.

The directive means that all interventional clinical trials, whether commercially funded or non-commercial, must meet these requirements. Amongst other things, every trial will need a sponsor. The EU Directive will have a number of implications for all parties involved in clinical trials. It enforces a number of new legal obligations for Chief/Principal Investigators and trial sponsors. In most commercially funded studies, the commercial company takes the role of 'sponsor' and is therefore responsible for the conduct of the study. The Directive has particular implications for studies funded by charities and studies funded by other non-commercial organisations. The original publication of the EU Clinical Trials Directive (2001/20/EC) can be downloaded directly from:

http://europa.eu.int/eur-lex/pri/en/oj/dat/2001/l_121/l_12120010501en00340044.pdf

Concerns have been raised about issues in these regulations including:

- identifying a trial sponsor
- delays to trial start up

- responsibilities
- levels of monitoring required

3.3 Regulatory Requirements and guidelines

Several more or less mandatory guidelines exist for clinical trials. In the following the most relevant regulations and guidance/guidelines regarding CDM process will be mentioned.

Although most of the described regulations apply primarily for drug admission, they can also be seen as guidelines for any trial sponsored by parties other than drug companies. Here we only want to give a very brief outline, a thorough review can be found in Section 21 - “Ethico-legal issues related to multicentric, post genomic Clinical Trials” in Part 2 of this document.

3.3.1 ICH – Good Clinical Practice

The International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) [ICH] was founded by regulatory agencies and experts from the pharmaceutical industry from Europe, Japan and the United States of America in 1991. The purpose of this initiative is to harmonize worldwide regulatory requirements for the registration of new human therapeutics. The aim is to reduce the need to duplicate the testing carried out during the research and development of new medicines by recommending ways to achieve greater harmonization in the interpretation and application of technical guidelines and requirements for product registration. Harmonization would lead to a more economical use of human, animal and material resources, and the elimination of unnecessary delay in the global development and availability of new medicines while maintaining safeguards on quality, safety, and efficacy, and regulatory obligations to protect public health [CHO2004], [MedDRA].

ICH has published a variety of guidelines that are divided into four major categories; ICH Topic Codes are assigned according to these categories:

Q: "Quality": related to chemical and pharmaceutical Quality Assurance.
Examples: Q1 Stability Testing, Q3 Impurity Testing

S: “Safety” related to in vitro and in vivo pre-clinical studies.
Examples: S1 Carcinogenicity Testing, S2 Genotoxicity Testing

E: "Efficacy" Topics, related to clinical studies in human subject.
Example: E6 Good Clinical Practices.

M: "Multidisciplinary” related to cross-cutting topics that do not fit uniquely into one of the above categories.
Examples: M1: Medical Terminology (MedDRA) [MedDRA].

Clinical trials compliant to the ICH guidelines are normally worldwide accepted, because ICH is a project of the most important regulatory agencies and pharmaceutical companies.

Most of the guidelines summarized under the topic “Efficacy” and two of the guidelines under the Topic “Multidisciplinary” are related to clinical data management practices. Four of them will now be described in more detail as these most affect clinical data management.

ICH Harmonized Tripartite Guideline General Considerations for Clinical Trials (E8) (ICH 1997)

This document is intended to give an overview of the ICH clinical and safety guidelines for readers who are unfamiliar with the ICH requirements.

The aim of this guideline is to describe internationally accepted principles and practices in the conduct of clinical trials and to facilitate the acceptance and evolution of foreign clinical trial data by promoting a common understanding of general principles, approaches and the definition of relevant terms

The guideline addresses different aspects related to planning the objectives, design, conduct, analysis and reporting of clinical trials [ICH], [RON2000].

ICH Harmonized Tripartite Guideline for Good Clinical Practice (E6 (R1)) (ICH 1996)

ICH E6 (R1) Guideline for Good Clinical Practice (ICH GCP) sets forth an international ethical and scientific quality standard for the conduct of clinical trials that involve the participation of human subjects. The GCP guideline covers all aspects of preparation, design, monitoring, recording, reporting and archiving of clinical trials. Copies of this guideline can be accessed directly from the following link:

http://www.ncehr-cnerh.org/english/gcp/.

The guideline includes an initial glossary of terms with harmonized definitions for terms used in clinical trials (e.g. adverse event, Case Report Form). These harmonized definitions should be used in all documents generated during the conduct of a clinical trial. The essential documents for the conduct of a clinical trial are defined and the responsibilities of the participating parties are clarified.

The ICH GCP guideline indicates that quality control should be applied to each stage of the data handling to ensure that all data are reliable and have been processed correctly. Any changes or corrections to a CRF should be dated, initialed, and explained (if necessary) and should not obscure the original entry. This applies to both, written or electronic changes or corrections. An audit trail should be maintained. The ICH GCP guideline suggests that certain quality control procedures should be employed during data management processing and the procedures should be available for audit [CHO2004], [ICH], [RON2000].

ICH Harmonized Tripartite Guideline Clinical Safety Data Management: Definitions and Standards for Expedited Reporting (E2A)

This guideline falls under the broad topic of Clinical Safety Data Management and is associated with topics E2B, E2C and M1. These guidelines provide standard definitions, an international medical terminology (via the Medical Dictionary for Drug Regulatory Affairs – MedDRA) and standard data elements for reporting medical information as well as timeframes for reporting safety information to regulatory authorities [ICH], [RON2000].

ICH Harmonized Tripartite Guideline Structure and Content of Clinical Study Reports (E3) (ICH 1995)

This guideline is intended to facilitate the compilation of a single worldwide core clinical study report acceptable to all regulatory authorities. By developing a report that is complete,

unambiguous and organized, it is hoped that the review of such reports by regulatory agencies will be made easier [ICH], [RON2000].

3.3.2 Rules / Guidelines Supporting ICH process

Rules and guidelines that support the ICH concepts but are not formally part of the ICH process are released from the US Food and drug administration and the European Commission. They are of interest to Clinical Data Management as they provide specific instructions to companies seeking compliance with the ICH requirement. The most important will be described in the following.

Directive of the European Parliament and of the Council of the EU

The first ever EC directive on Clinical Trials was "DIRECTIVE 2001/20/EC", which was released in 2001. This directive is officially entitled „On the approximation of the laws, regulations and administrative provisions of the Member States relating to the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use" [DIR2001].

2005 it was supplemented by "COMMISSION DIRECTIVE 2005/28/EC**"** entitled

"Laying down principles and detailed guidelines for good clinical practice as regards investigational medicinal products for human use, as well as the requirements for authorization of the manufacturing or importation of such products" [DIR2005].

The purpose of both directives is to harmonize the regulation of clinical trials in Europe since requirements set down in ICH guidelines are not binding in countries of the European Union. The directives also aim among other things to make legal the GCP inspections performed by regulatory authorities in organizations participating at clinical trials. The main statement is that ICH GCP will have to be followed and all parties conducting trials will now be open to inspection. The directives reinforce those ethical principles which have been already harmonized between member states and also propose some new standards.

The main topics of the directive from 2001 are Protection of Trial Subjects, Ethics Committee Opinion, Commencement of a Clinical Trial, and Conduct of a Clinical Trial

Exchange of Information, Manufacture and Import of Investigational Medicinal, Compliance with Good Clinical Practice and Notification of Adverse Events (SAE: Severe Adverse Events) and Adverse Reactions (SUSAR: Suspected Unexpected Severe Adverse Reactions) [RON2000].

FDA (Food and Drug administration)

The FDA (Food and drug administration) is the US agency for admission of drugs.

The regulations set by the FDA are also important for European studies, because of the great importance of the US market.

The FDA-derived documents relate to the proposal from the agency to move from paper regulatory submission to electronic submissions in stages. Since these topics represent rapidly growing areas, the guidance is updated periodically. The FDA's goal is to establish an approach for submitting electronic applications that create minimal work and reduced costs,

as well as encouraging consistency in information transfer requirements across agency [FDA], [RON2000].

21 CFR part 11 – Electronic Records, electronic Signatures

21 CFR part 11 describes the criteria under which the FDA will consider electronic records and signatures to be generally equivalent to paper records and handwritten signatures. It applies to any records required by the FDA or submitted to the FDA under agency regulations.

To reinforce Part 11 compliance, FDA has published a compliance policy guide and in addition numerous draft guidance documents to assist the sponsor for Part 11 compliance [FDA], [RON2000].

21 CFR part 312 – Investigational new Drug Application and 21 CFR part 314 – Applications for Food and Drug Administration Approval to market a new Drug

21 CFR Parts 312 and 314 contain regulations for submitting a new drug application to the FDA. Both regulations have direct relevance to CDM, especially the handling of CRFs. They also cover requirements for investigator's record-keeping and record retention. They pertain to the inclusion of CRFs and tabulations in an application [CHO2004], [RON2000].

FDA Guidance for Industry – Computerized Systems Used in Clinical Trials

This guidance provides general principles that are to be followed when computerized systems are used to create, modify, maintain, archive, retrieve, or transmit clinical data intended for submission to the agency. The addressed requirements are similar to those described in 21 Part 11 for electronic records and electronic signatures [FDA], [RON2000].

3.4 Consequences for Clinical trials

Compliance with the above mentioned standards provides public assurance that the rights, safety and well-being of trial subjects are protected, in a way consistent with the principles that have their origin in the Declaration of Helsinki, and that the clinical data are credible.

The following principles have to be strictly followed:

1. Clinical trials should be conducted in accordance with the ethical principles that have their origin in the Declaration of Helsinki, and that are consistent with GCP and the applicable regulatory requirements
2. Before a trial is initiated, foreseeable risks and inconveniences should be weighed against the anticipated benefit for the individual trial subject and society. A trial should be initiated and conducted only if the anticipated benefits justify the risks
3. The rights, safety, and well-being of the trial subjects are the most important considerations and should prevail over the interests of science and society
4. The available non-clinical and clinical information on an investigational product should be adequate to support the proposed clinical trial
5. Clinical trials should be scientifically sound, and described in a clear, detailed protocol
6. A trial should be conducted in compliance with the protocol that has received prior ethics committee approval

7. The medical care given to, and medical decisions made on behalf of subjects should be the responsibility of a qualified physician
8. Each individual involved in conducting a trial should be qualified by education, training and expertise to perform his or her respective tasks
9. Freely given informed consent should be obtained from every trial participant prior to clinical trial participation
10. All clinical trial information should be recorded, handled, and stored in a way that allows its accurate reporting, interpretation and verification
11. The confidentiality of records that could identify subjects should be protected, respecting the privacy and confidentiality rules in accordance with the applicable regulatory requirements.

There is a variation between different countries in Europe in interpretation of the directive and the common problems encountered by established academic (non-commercial) trial groups in maintaining and opening new multinational clinical trials for cancer.

The practical application of these principles requires that clinical trials have distinct components built into them. These include:

1. **Relevant and appropriate study rationale**
   A study rationale and motivation which does not ask relevant and important questions is unethical. Whilst maintaining the highest standards of clinical research it is important that clinical trials are based on priority research questions. Relevant and important questions should be problems that significantly affect local and regional population. Study rationale should demonstrate that the study question has not been substantially answered and that adequate systematic review of the subject under discussion was done.

2. **Optimal study design**
   Appropriate designs are critical in contributing to answering scientific questions. The design must therefore demonstrate a high probability for providing answers to specific research questions. Adequate supporting information and explanation on the study sample size and study population must also be provided.

3. **Investigator competence**
   The investigator's competence is assessed by two major parameters: Technical and humanistic. Technical competence which includes research competence is assessed by education, knowledge, certification and experience. Humanistic parameters require compassion and empathy. This is provided by a proper clinical and research environment.

4. **Balance of risks and benefits for participants**
   A risk benefit analysis of the study should precede the conduct of the research itself. Risk-benefit analysis should take full cognisance of benefits and harms beyond the life of the study itself, particularly in the case of chronic life threatening conditions. Alternative ways of providing benefits to the patients might be available without research; thus the distinction between the probability of harm and the possible benefits of the effects must be made.

5. **Transparency**
   A clinical trial in the European Community can only obtain approval, if the trial information is registered in the EudraCT database (see 2.3.3.1). The database will serve to promote transparency to prevent unnecessary trials.

6. **Patient privacy**
   The patient's privacy, data protection and security are essential tasks in clinical trials. In ACGT Workpackages 10 and 11 are dealing with these issues.

7. **Ethics**
   In respect to ethics clinical trials have to pass ethical committees and regulatory or legal authorities. In ongoing trials data and safety monitoring committees are included.
   - Ethical Committees. They are usually made up of lawyers, medical practitioners, bio-ethicists and community representatives
   - Data and Safety Monitoring Committees (DSMC). These committees oversee ongoing clinical trials with respect to treatment, efficacy and safety. In the advent of clear evidence of efficacy or harm, prior to the end of the trial, premature termination can be recommended on ethical grounds
   - The Regulatory Authority which is responsible for reviewing the study design

8. **Impartial oversight of consent procedures**
   Informed consent is a necessary but not sufficient requirement for ethical conduct. Obtaining informed consent implies the provision of information to potential participants regarding the nature of the research procedure, scientific purpose and alternatives to study participation. Participants' comprehension is addressed by laying out this information in a clear and simple style, including the use of the participant's language. The conditions under which the consent is granted must be free of coercion, undue influence or incentives. Treatment for a given condition, which might be an attribute of the clinical trial design, should not be denied by the refusal to participate. Withdrawal from the clinical trial at any time will not result in undue clinical penalties to the participant.

9. **Safety monitoring**
   Safety monitoring of participants during and for defined periods after a clinical trial is an ethical requirement. This involves the prompt reporting of serious adverse events and the appropriate management of such an event. (SAE, SUSAR, see 2.3.4.2)

3.5 Current Status of clinical research by trial group and country

3.5.1 Questionnaire about implementation of the EU CTD

A brief questionnaire about implementation of the EU CTD and its impact on newly opening paediatric trials was carried out by Kathy Pritchard-Jones, UK (personal communication, 2006 [PRI2006]). The questionnaire was distributed to:

- Chief Investigators of major European Clinical Trials
- National childhood cancer societies of each European Country, or
- Senior paediatric oncologist to represent national activity, if no group

An aim of the questionnaire was to survey the practical solutions found and the on-going problems in each of the following areas:

- Sponsorship
- Definition of clinical trial
- Insurance/indemnity
- Pharmacovigilance
- Time to open trials

- Other comments

Responses had been received from the majority of countries and some international clinical trial groups. Some findings are summarised in the following tables

➲ Have you found a national solution to sponsorship of non-commercial trials?

Scope of solution	Nature of solution	Countries
YES, Generic	NHS hospitals linked to UKCCSG & leukaemia trials	UK
	National Association of Paediatric Oncology (AIEOP) – *national law allows scientific society to sponsor not for profit, non-commercial trials*	Italy
NOT REALLY, Trial-specific	University or Cancer centre employing chief investigator has accepted sponsor role	France
	Variation in willingness to accept this role	Germany, Czech Republic, Lithuania
	Investigator-led	Denmark, Spain, Finland, Austria, Norway
NO solution yet identified	No solution	Belgium
	Not yet required for new member states	Romania
	Not yet solved	Turkey
	Application in process	Lithuania, Hungary

➲ Does your country require a single EU sponsor and how does it deal with the issue of responsibility across national boundaries?

Country		Comments
Hungary, Czech Republic, Austria, Belgium	**Yes**	EORTC trials provide sponsorship
Denmark	**?Yes**	Communications between CI in Denmark and Danish Medicines Agency require authorisation letter from main sponsor
Norway	**Yes**	Communications between CI in Norway and Norwegian authorities require authorisation letter from main sponsor
Germany	**Yes**	Letter to German Chancellor
Lithuania	**Yes**	Responsible person authorised by sponsor can operate in Lithuania
France, Ireland, Italy, Spain, Finland, United Kingdom, Romania	**No**	
Turkey	**N/A**	Not yet relevant

➲ Does the EU CTD apply in your country to 'diagnostic' trials where the 'intervention' is to test a risk stratification based on a biological or imaging tests for their association with response and outcome?

Country	Comments
Czech Republic	Uncertain
Denmark, Norway	No
France	If trial has 'standard arm' that = best accepted practice over many years and no stopping rule and not in randomised comparison, then falls outside definition of clinical trial but requires separate protocol/recommendation
Germany	Yes, only epidemiological studies excluded
Italy	Unsure, lack of clarity of what is 'non-interventional'
Romania	?
Turkey	Diagnostic trials do not require sponsor, just ethical approval and consent
United Kingdom, Finland, Spain, Belgium, Austria	Yes

Sponsorship remains a major issue for the majority of countries where there is still no national solution to sponsorship requirements for academic clinical trials. Sponsorship tends to be provided by the academic institution of the lead investigator. The issue of the need for a pan-European sponsor is particularly problematic for Belgium, where the regulatory authorities do not recognise the existence in law of a "national sponsor" who can assume delegated sponsor responsibilities. In the Scandinavian countries, sponsor responsibilities can be delegated to a national representative, but the latter has to have written permission from the sponsor in order for them to communicate trial issues to their national regulatory authorities. Encouragingly several countries have found a "workable" solution in this area, even though there is still lack of clarity about the need for and definition of a 'single European sponsor'.

Regarding the definition of a clinical trial, most countries have not succeeded in having any other than epidemiological studies escape the EU CTD. France has found a novel solution to reduce the administrative workload of opening the new European rhabdomyosarcoma protocol, EpSSG RMS 2005. They have split this into two protocols, one of which describes the use of IVA chemotherapy for rhabdomyosarcoma, which is accepted standard practice. As such, any patients following these arms would not be considered to be in a clinical trial and therefore would not require reporting of adverse events etc.

Indemnity Insurance Issues

There continues to be large variations in the requirement for 'no fault' insurance and unaffordable variations in premiums. Several countries have made progress in this area. The German Cancer Society has been involved in definition of ten risk levels with a simplified and much more affordable insurance level. The Italian Parliament has legislated in December 2005 that academic trials do not require 'no fault' insurance or indeed any form of insurance over and above that which is supplied by normal hospital indemnity.

Ethical Approval

Although the EU CTD requires a single ethical approval for clinical trials with defined timelines, there continues to be some variation in how this is interpreted. In many countries, although there is notionally a single national ethical approval process, the requirement for local ethical committees to decide whether or not a trial can be run in a particular institution, is still causing additional bureaucracy and delays. Most regulatory authorities are complying with the timelines of a 60-day turnaround for approval. The majority of delays are appearing due to the overall burden of bureaucracy. Time to open new studies since May 2004 has ranged from 3 months for a new drug trial that could be fast-tracked in France to several months to years for the majority and an answer of "indefinite" from the Belgians due to their current sponsorship difficulties.

3.6 References

[WFMC] Workflow Management Coalition, http://www.wfmc.org/

[RAN2003] Rang HP, Dale MM, Ritter JM, Moore PK (2003). *Pharmacology* 5 ed. Edinburgh: Churchill Livingstone. ISBN 0-443-07145-4

[FIN1999] Finn R, (1999). "Cancer Clinical Trials: Experimental Treatments and How They Can Help You." Sebastopol: O'Reilly & Associates. ISBN 1-56592-566-1

[CDISC] Clinical Data Interchange Standards Consortium. http://www.cdisc.org, last accessed: 24.04.06

[CHO2004] Chow S, Liu P (2004). Design and Analysis of Clinical Trials. Second Edition, JohnWiley & Sons, Hoboken, New Jersey.

[DIR2001] DIRECTIVE 2001/20/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL (2001). http://eudract.emea.eu.int/docs/Dir2001-20_en.pdf. last accessed: 24.04.2006

[DIR2005] COMMISSION DIRECTIVE 2005/28/EC (2005). http://pharmacos.eudra.org/F2/eudralex/vol-1/DIR_2005_28/DIR_2005_28_EN.pdf. last accessed: 24.04.2006

[FDA] FDA Food and Drug Administration. http://www.fda.gov/. last accessed: 24.04.2006

[ICH] ICH International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. www.ich.org. last accessed: 24.04.2006

[KUS2003] Kush R (2003). The world of Standards for Clinical Research, http://www.touchbriefings.com/pdf/16/Kush.pdf, last accessed: 24.04.2006

[MedDRA] MedDRA and the MSSO. http://www.meddramsso.com/MSSOWeb/index.htm, last accessed: 24.04.2006

[RON2000] Rondel R, Varley S, Webb C (2000). Clinical Data Management. 2. Aufl, John Wiley & Sons, Chichester.

[WEB2004] Weber, Ralf (2004). Terminologiebasierte Erstellung von rechnerunterstützten Dokumentationssystemen in klinischen Studien; Dissertation Medizinische Fakultät

Heidelberg.

[GAL2005] Galea-Lauri, Joanna; Forster, Louise (2005). An overall Guide to Interventional Clinical Trials of Medicinal Products for Researchers at ICH/GOSH [GOSH ICH/05/S03/02 (Revised 18th March 2005)]. Approved by: Professor David Goldblatt (Clinical Director, Research and Development) http://www.ich.ucl.ac.uk/ich/r&d/ctguidelines.pdf, last accessed: 21.04.2006

[PRI2006] Pritchard-Jones, Kathy (2006). A brief questionaire about the implementation of the EU CTD and its impact on newly opening paediatric trials, personal communication.

4 The ACGT Clinical Studies

This section provides a short description of the Clinical Studies (Trials) that have been selected for demonstrating the ACGT infrastructure. Please refer to WP12 and D12.1 fro a detailed description of the studies, their rational and objectives, and other details.

4.1 The ACGT – TOP Study on Breast Cancer

4.1.1 Breast Cancer

Breast cancer is the *commonest cancer in women in the world,* in both industrialized and developing countries. Over a million women will be diagnosed with breast cancer worldwide in 2004[1]. More than 40,000 women will die this year of metastatic breast cancer in the United States alone and more than 200,000 new cases of cancer will be detected[2]. The mortality rate around the world, especially in developing countries, is much higher, making breast cancer a *significant public health problem*.

Much progress has been made over the past decades in our understanding of the epidemiology, clinical course and basic biology of breast cancer. *Identified risk factors* include:

- *Family history (genetics).* Identified gene mutations represent a tiny fraction of all breast cancers, much less than 10% overall. But if present, they confer considerable life-time risk compared to the general population.

- *Reproductive and hormonal life*, e.g., early menarche, no pregnancy or late age at first birth, late menopause, hormonal factors such as high levels of free oestrogen, long-term use of oral contraceptives or menopausal hormone replacement, or other factors that increase life-time exposure to oestrogen.

- *Lifestyle, particularly diet* and *exposures* to carcinogenic agents.

Breast cancer is both genetically and histopathologically *heterogeneous*, and the mechanisms underling breast cancer development remains largely unknown. The genetic background of patients and the tumour's genetic and epigenetic anomalies create, in combination, molecularly distinct subtypes arising from distinct cell types within the ductal epithelium. This genetic complexity underlies the clinical heterogeneity of breast cancer limiting a rational selection of treatment tailored to individual patient/tumour characteristics, thus breast cancer patients diagnosed with the same stage of disease often have remarkably different responses to therapy and overall outcome.

However, risk stratification based on these guidelines is far from perfect, and much progress is needed to identify those patients who really need adjuvant systemic therapy. Several independent groups have conducted comprehensive gene expression profiling studies with the hope of improving upon traditional prognostic markers used in the clinic.

[1] Pisani P., Bray F, Parkin DM. Estimates of the worldwide prevalence of cancer for 25 sites in the adult population. Int J Cancer 2002;97:72-81

[2] Jemal A, Murray T, Samules A, et al. Cancer Statistics,2003. Cancer J Clin 2003;53:5-26

In current clinical practice, the majority of patients with early breast cancer will receive some form of systemic adjuvant therapy (chemo- and/or endocrine therapy). Clinical parameters, such as lymph node status, tumor size and histological grade can provide prognostic information, and are summarized in clinical guidelines, such as the National Institutes of Health (NIH) or the St Gallen consensus criteria, in order to assist clinicians and patients in adjuvant therapy decision-making. The analysis of additional single molecular markers can also contribute to the therapeutic decision making. Although all of these factors have been correlated to patients' survival in general, the same prognostic profile often results in dissimilar clinical outcomes in individual patients.

Thus, conventional prognostic factors provide insufficient information to evaluate the heterogeneity of this disease and to make treatment more effective for individual patients. One problem faced by present cancer therapy is the over-treatment of patients with chemotherapy, which is associated with severe toxicity and increasing healthcare spending without clear survival benefit over untreated controls. Because of the lack of adequate predictive markers, nearly all patients receive routinely standard treatment in spite of grim changes of deriving any benefit. Therefore, the identification of molecular markers predictive of patients' responsiveness to treatment is becoming a central focus of translational research.

4.1.2 Objectives of the ACGT-TOP study

Although breast cancer mortality has declined in the last two decades, breast cancer continues to represent a major threat to the lives and productivity of women. The number of effective treatments for breast cancer rise; however, the benefit from specific treatments to individual patients and the adverse events they experience vary considerably. Efficacy and safety of anticancer therapies may depend on tumour, treatment, and host characteristics.

Small variants in the germline DNA sequence (genotype) may lead to different expression of the encoded protein or to the expression of altered protein, and thus to a different health outcome (phenotype). Most genetic variants occur in noncoding regions of the genome, and although such variants may result in functional consequences, most known variants that are associated with clinically important functional change are in the exons that code for protein expression. While the clinical importance of a large number of pharmacogenetic variants is becoming clearer, the significance of the majority remains speculative while we await larger trials. Preliminary pharmacogenetic data strongly suggest an important role for the use of germline genetic information in the individualization of treatment and prevention of breast cancer. The potential value of these data as individual genotypic predictors may be valuable or, more likely, patterns of genetic markers analogous to the expression profiles obtained from tumour tissue may allow more powerful prediction of who will respond best to a specific treatment or regimen. For the potential of genomic research to be fulfilled, prospective trials with clinical outcomes as end points will have to include the collection of germline DNA. Pharmacogenetics may play a significant role in several aspects of breast cancer including prognosis at the time of diagnosis, response to specific treatments, and likelihood of adverse events to specific treatments.

Once a decision is made to administer systemic therapy, only a handful of genes or proteins are used to select specific treatments for breast cancer patients. Early results suggest that patterns of gene expression determined on primary tumours may predict sensitivity or resistance to common breast cancer treatments. A central, but often unrecognized goal of pharmacogenetic research is to use the revolution in genomics to allow the benefit of treatment and avoidance of toxicity to be made available to all patients, rather than only the subgroup that can tolerate currently used therapeutic regimens and respond well to them.

Knowledge of the likelihood of response to treatment and the predictability of side effects may assist in individualizing treatment for women diagnosed with breast cancer.

DNA Microarrays are a versatile technology that can be used for both patient genotyping and tumour gene expression profiling. Thus provide a high-throughput screening tool for the identification of molecular patterns that associate significantly with particular clinical and pharmacological behaviours.

Identifying “high-risk” patients who clearly need systemic adjuvant therapy is not good enough: we still do not know which type of therapy will be most efficient for each individual patient. Identifying markers that can reliably predict response to particular drugs remains a great challenge. To this end, the neoadjuvant approach is very attractive, as it provides an *in vivo* assessment of treatment sensitivity. Several studies have already used a genome-wide approach to identify gene expression profiles that correlate with chemo- or hormonosensitivity.

The ACGT TOP study aims to identify biological markers associated with pathological complete response to anthracycline therapy (epirubicin), one of the most active drugs used in breast cancer treatment. To this end, the neoadjuvant approach is very attractive, as it provides an *in vivo* assessment of treatment sensitivity without affecting adversely survival. Tumor samples drawn at the time of pre-treatment biopsy will be frozen and used to perform oligonucleotide based microarrays (Affymetrix). This technique allows the evaluation of thousands of genes and ultimately provides us with the tumor genetic profile. Homogeneous genetic profiles (genetic clusters) that might be identified, will be correlated with the efficacy of single-agent epirubicin. This correlation will allow us to address the secondary end-point of this study, which is the identification of other genes or eventually a genetic profile playing a role in the determination of sensitivity to anthracyclines. The study target two patient groups:

4.1.2.1 Early beast cancer

The concept of delivering chemotherapy as primary treatment in early breast cancer patients is attractive because chemosensitivity of the tumor can be assessed "in vivo" allowing for a more "tailored" approach in systemic therapy. This treatment strategy can support further therapy with non-cross resistant agents in cases showing moderate or poor sensitivity to primary chemotherapy. This latter point is particularly attractive because potentially effective “salvage treatments”, such as the taxanes, are nowadays available.

Large international phase II and, more importantly, phase III trials have demonstrated that three to four cycles of primary chemotherapy are feasible and do not compromise either the efficacy of loco-regional treatments (surgery and radiotherapy) or long-term survival. Moreover, tumor down-staging achieved after primary chemotherapy can lead to breast-conserving surgery in those patients with large operable tumors, who would otherwise be candidates for mastectomy. The potential advantages of primary medical therapy, and the increasing use of adjuvant chemotherapy in node-negative patients, support the design of a study in which early breast cancer patients will be treated with primary chemotherapy.

The most common and acute dose-limiting hematological toxicity seen with adriamycin and epirubicin is reversible leucopenia and/or neutropenia, although anemia and thrombocytopenia can also occur. Non-hematological toxicities include: alopecia, nausea and vomiting, diarrhea and stomatitis, and cutaneous and hypersensitivity reactions. All these toxicities are acute, reversible and usually manageable, particularly with the advent of new anti-emetic drugs. Of greater concern are two possible long-term toxicities, namely secondary leukemia and cardiotoxicity.

Supported by "in-vitro" and preliminary "in-vivo" data, this study is designed to test prospectively the value of topo II alpha gene amplification and protein overexpression in predicting the efficacy of anthracyclines. To our knowledge this is the only prospective trial worldwide which is attempting to prospectively clarify the predictive value of this interesting biological marker. This study could have important practical implications in the daily clinical management of early breast cancer patients because, if the trial confirms that topo II α gene amplification and/or protein overexpression are associated with high efficacy of anthracyclines, while topo II α normal/deleted gene and low protein content are associated with modest efficacy, an important step forward in the direction of anthracycline "tailoring" would be accomplished.

The practical advantage of this approach would be to use anthracyclines primarily in patients who are supposed to derive the largest benefit, thus sparing the long-term anthracycline-related toxicity (i.e. secondary acute myeloid leukemia, cardiac dysfunction, and amenorrhea/sterility in case of fertile women) to those patients for whom no significant gain in antitumor activity is anticipated.

4.1.2.2 Inflammatory and locally advanced breast cancer

Inflammatory breast cancer, perhaps the most aggressive form of breast neoplasia represents 1 to 3% of newly diagnosed breast malignancies. The entity is diagnosed on clinical grounds, based on the presence of erythema and edema (peau d'orange) of the skin of the breast, as well as ridging. Most inflammatory cancers present as diffuse infiltration of the breast without a well-defined tumor. Dermal lymphatic invasion is present in most patients, but this feature is not a necessary component of the diagnosis. Most inflammatory breast cancers are poorly differentiated ductal carcinomas and are ER- and PR- negative.

Locally advanced breast cancer (LABC) encompasses a heterogeneous group of patients including those with neglected slow growing tumors as well as those with biologically aggressive disease. Locally advanced breast cancer is a relatively uncommon presentation in the economically developed world accounting for 5 % of cases in major centers. However, in medically underserved area and in many countries, LABC represents 30 to 50% of newly diagnosed breast cancers**.**

Inflammatory and locally advanced breast carcinomas are both associated with poor prognosis. With surgery and/or radiotherapy alone, the prognosis in LABC is very poor. This poor long-term outcome prompted the introduction of primary chemotherapy or hormonotherapy, with the first reports published in the 70's. Such a multimodality approach has led to a significant improvement in LABC outcome. Clinical complete remissions were reported in 10 to 20% of patients treated in this manner in most clinical trials. However, only two thirds of the patients with a clinical complete response are found to have a pathologic complete response. Several authors have demonstrated that the achievement of a pathologic complete response is an excellent predictor of long-term survival. With standard anthracycline-based regimens, pathologic complete response rates range from 3.5% to 12%. The addition of taxanes to anthracyclines in the neo-adjuvant regimen significantly increases pathologic complete response and improves survival of patients who achieve pathologic complete response.

Dose intensity can be increased either by dose escalation and/or by reducing the interval between the cycles. Interest in dose-intensity is based on the observation that, in experimental models, a given dose kills a certain fraction rather than a certain number of exponentially growing cancer cells. Results of dose intensification by increasing dose of chemotherapy (high dose chemotherapy with stem-cell support) have been disappointing in the adjuvant setting.

More recently, the dose density hypothesis, which refers to the administration of drugs with a shortened inter-treatment interval, has been tested. It has been hypothesized that a more frequent administration of cytotoxic therapy could be a more effective way of minimizing residual tumor burden than dose escalation.

In LABC, a relatively small trial compared a conventionally dosed neo-adjuvant regimen to a dose-dense regimen. The study failed to show an improvement in disease free survival (DFS) with the dose dense combination. However, it is interesting to note that, with a median follow-up of 5 years, the short dose-dense regimen was as effective as the longer CEF treatment, with no increased rate of cardiotoxicity or leukemia.

In the present study, we plan to use a dose-dense administration of epirubicin (100 mg/m^2/2 weeks). We keep the same drug as for early breast cancer but we use a slightly more aggressive regimen with a higher dose-density. The feasibility of the administration of epirubicin 100 mg/m^2 every two weeks with granulocyte-growth factor support has been shown in the neoadjuvant, metastatic and adjuvant settings with acceptable toxicity. This neoadjuvant epirubicin regimen may be completed by adjuvant chemotherapy, such as taxane-based regimens, since the sequential approach (anthracyclines $\rightarrow$ taxanes) has been suggested superior to anthracyclines regimen in LABC.

4.2 The ACGT Paediatric Nephroblastoma Study

4.2.1 Paediatric Nephroblastoma or Wilm's tumour

Although rare, Wilms' tumour is the most common primary renal malignancy in children and is associated with a number of congenital anomalies and documented syndromes. Appropriate laboratory, radiologic and pathologic investigations are necessary for accurate diagnosis and subsequent staging; information which is essential to generate a multidisciplinary treatment plan utilizing surgery, chemotherapy and radiotherapy.

Wilms' tumour was the first solid malignancy in which the value of adjuvant chemotherapy was established. Multimodality treatment has resulted in a significant improvement in outcome from approximately 30% in the 1930s to more than 85% in the modern era. Although the National Wilms' Tumour Study Group in North America and the International Society of Paediatric Oncology, involving European and other countries, differ philosophically regarding the merits of preoperative chemotherapy, outcomes of patients treated with either up-front nephrectomy or preoperative chemotherapy have been excellent.

The results that have been achieved in children with Wilms' tumours support the strong value of the multidisciplinary team approach to cancer. The goal of current clinical trials is to reduce therapy for children with low-risk tumours, thereby avoiding acute and long-term toxicities. Challenges remain in identifying novel molecular, histological and clinical risk factors for stratification of treatment intensity. This could allow a safe reduction in therapy for patients known to have an excellent chance of cure with the current therapy, while identifying, at diagnosis, the minority of children at risk of relapse, who will necessitate more aggressive treatments.

4.2.2 Rationale and Objectives of the Nephroblastoma study

The ACGT Nephroblastoma clinical study is based on SIOP 2001 study, a continuation of the philosophy of the former SIOP studies. The basic idea has always been: Collect a lot of reliable data by working together on an international base and answer questions which can

be of direct importance for the outcome of the patients. SIOP 2001 is based on the results of the previous SIOP trials and studies as well as on the results of the NWTS protocols and its specific objectives are:

- To adapt therapy to the known individual risk of the patient and increase survival for blastemal predominant tumours after preoperative chemotherapy by intensifying therapy and minimise acute and late toxicity without jeopardising event free survival and survival by reducing treatment for patients with focal anaplasia, for stage I patients with intermediate risk tumours, and for stage II and III patients with intermediate risk tumours by randomising doxorubicin.

- To test the treatment hypothesis that doxorubicin is not necessary in patients with intermediate risk tumours and local stage II or III by a multicentric prospective randomised trial.

- To prospectively analyse different histological components of nephroblastoma with a special emphasis on a percentage of blastemal component which might be of prognostic significance

- To reduce the number of drug administrations, hospital visits and thereby costs in the preoperative phase

- To collect material for performing biological studies with specific aims and clinical research scenarios.

Besides the excellent prognosis of children with Wilms tumour there is a well known risk of unnecessarily administered chemotherapy by treating children preoperatively without histologically proven diagnosis. This risk could be abolished by finding a specific marker for Wilms tumour in serum, which is lacking today. Immunogenic tumour-associated antigens have been reported for a variety of malignant tumours including brain tumours, prostate, lung and colon cancer. The purpose of the ACGT nephroblastoma study is to find such a marker by searching for a pathognomonic antigen pattern in patients with Wilms tumour. Serum from a specific patient will be tested against newly identified Wilms tumour antigens. As a result in each patient there will be a specific pattern of antigens found. This pattern will be correlated to the histological subtype of the tumour, the gene expression profiling of the tumour, the response to chemotherapy and the outcome of the patient. The study is described in further detail as Scenario S2 in section 6 of this document.

4.3 In Silico Modeling of Tumor Growth

The aim of the third ACGT study is to provide clinicians with a decision support tool able to simulate within defined reliability limits the response of a solid tumour to therapeutic interventions based on the individual patient's data. An intermediate goal of the study would be to provide researchers with a versatile platform for integrating experimental and clinical knowledge and performing exploratory experiments *in silico.*

The *In Silico* Oncology clinical research will be based on the two other clinical trials incorporated in ACGT (the expansions of the nephroblastoma SIOP 2001/GPOH and breast cancer TOP trial), and would aim at developing, optimizing, validating and clinically adapting a computational system, denoted by the specially coined term "Oncosimulator" that would serve as simulation model of tumour response to chemotherapy. The most critical biological phenomena (e.g. metabolism, cell cycling, geometrical growth or shrinkage of the tumour,

cell survival following irradiation or chemotherapeutic treatment, necrosis, apoptosis etc.) will be thus spatiotemporally simulated using a variety of clinical, radiobiological, pharmacodynamic, molecular and imaging data.

Furthermore, Virtual Environments designed to represent 3D (and to some degree also 4D) data and to provide intuitive interactive methods to explore this data will be applied for *the virtual reality visualisation* of both medical images and *in silico* oncology simulation results. The objective to "involve" the researcher more, and bring her/him closer to her/his data in an effort to detect patterns and structures using the researcher's experience, expertise and cognitive abilities.

These innovative computational platforms definitely do not intend to replace the physicians' input but to add the possibility to investigate the impact of specific treatment-induced perturbations over several orders of magnitude – which currently is impossible with conventional imaging methods alone.

4.4 References

[FAN2006] Fan et al, Concordance among Gene-expression-based predictors for breast cancer. NEJM 355:560-566, 2006

[SHA2006] O'Shaughnessy: Molecular signatures predict outcome of breast cancer. NEJM 355:615-617, 2006

[PAI2006a] Paik S et al.: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. NEJM 351:2817-2826, 2006

[PAI2006b] Paik et al.: Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. J Clin Oncol (in press)

5 The ACGT Scenarios

5.1 Introduction

As discussed in Chapter 2, a scenario-based and user-focused methodology has been adopted for the engineering of user and system requirements in the project. The methodology is built upon object-oriented methodologies, domain modelling strategies, and scenario-based techniques to provide an analysis process for mapping user and application requirements to required services of the ACGT architecture.

Scenario descriptions **integrate what and how** the user carries out activities, with what thoughts and experiences accompany that activity. Scenarios facilitate concurrent design and evaluation by providing a common language. The methodology begins with Domain Analysis, which forms the foundation for Application/Systems Analysis and System Implementation. Representative end-users were included as active participants during each stage of the process and quickly became valued members of the technology team. This user-centred approach allowed all project team members to fully understand the operating environment for software and system development.

The methodology ensures that all development activities will be well focused and that resulting products are designed to meet '*real-world*' needs. It is worth mentioning at this point that the aim of the ACGT infrastructure is to facilitate clinical research and trials. As it is impossible to give an explicit and complete list of analytical tasks conducted in clinical trials (new tools and technologies are constantly developed, tools are used in various combinations defining “meta-tools”, etc…), the present Chapter lists a series of typical questions and problems to be addressed in the context of the analysis of clinical data. The explicit minimal list of services needed to achieve the mentioned goals will be iteratively refined.

A historical examination into the IEEE Standard Glossary of Software Engineering Terminology reveals an increasing awareness of the iterative nature of requirements development. In the 1983 glossary, “requirements analysis” is defined as “the process of studying user needs to arrive at a definition of system requirements” [IEEE 83]. This implies a one time, up front requirements definition activity. In the 1990 glossary, however, a second definition has been added for requirements analysis: “the process of studying and refining system, hardware, or software requirements” [IEEE 90]. This implies retrospective examinations of requirements with refinement steps, i.e., an iterative requirements engineering process.

Requirements are not completely known at the start of a system’s development. They cannot be specified completely up front in one voluminous document, but rather will evolve during the analysis phases of a project and beyond. The communities involved in the elicitation, including users, developers, and customers, all learn and grow during the system’s development and maintenance. This increasing knowledge possessed by the elicitation communities regarding the system should be utilized to improve the system, rather than prohibited because the requirements are to remain static.

> *In enabling this iterative process of capturing and refining user requirements, the “User Requirements” activity of the project continues for the full duration of the project. As a result the present document will be updated at regular intervals to*

(a) capture additional user requirements resulting from additional "post-genomic" scenarios and (b) advances in relevant technologies.

Based on such a methodology several indicative future scenarios have been developed. The scenarios presented can be separated into two main groups:

(a) user driven scenarios, representing the real needs of the ACGT users as understood at this point in time and expressed in an adequate level of details

(b) technology driven scenarios which are based exclusively on the "anticipated user needs" as understood by technology developers based on their experience and knowledge on the state-of-the art in the domain. These scenarios utilise published data which allows circumventing possible legal problems in sharing the data collected (or to be so) in the context of the "actual" ACGT scenarios (TOP, SIOP, etc…). The number of these technology driven scenarios is currently limited to three but this number is expected to increase in later revisions of this document.

The following two sub-sections provide a list of the key clinical questions that drive the creation of post-genomic clinical trials and a corresponding survey of the various categories of problems typically encountered in the analysis of clinical data, as summarized at various internal working meeting and as came out from a survey in the ACGT community.

Subsequently, the remaining sections describe the scenarios, providing for each: a reference to the published work if it exists, a description of the raw data, a survey of the methodology and of the results. They also provide a more or less ranked wish list of functionalities that should be made available in the ACGT platform. Finally, possible extensions to the present list of scenarios such that a broader set of functionalities of the ACGT platform is covered, is presented

5.1.1 List of key clinical questions

- ***Question 1 (class prediction):***
  Are there gene expression features that can discriminate/diagnose diseases or subtypes?
- ***Question 2 (class discovery):***
  What new disease subclasses are observed using gene expression profiles and what are the clinical implications?
- ***Question 3:***
  What are the genotypes that correlate to the responsiveness to therapy?
- ***Question 4:***
  Are gene expression profiles due to chromosomal alterations or to regulatory states?
- ***Question 5:***
  Can we define new diagnostic and classification entities that combine gene expression profiles and prior molecular biology knowledge?
- ***Question 6:***
  What are the molecular pathways or developmental states that are associated to the disease?
- ***Question 7:***
  How do results compare with previously published findings?

- ***Question 8:***

  Are there gene expression features that correlate with the stage of disease including lymph node involvement and metastasis?

- ***Question 8:***

  Are there gene expression features that predict side effects of chemotherapy or irradiation?

5.1.2 List of generic scenarios

Based on the State of the Art review done a number of scenarios, relevant to clinical trial research on cancer and the need to integrate and analyze multilevel biomedical data have been identified. Such scenarios are:

➲ ***Gene expression molecular signatures***

Identify from global gene expression the minimum set of genes whose expression levels, also referred to as molecular signatures, associates with significance to a particular trait (clinical or phenotypic).

➲ ***Molecular classification of tumours***

Use global gene expression of a set of tumour specimens in order to identify new subclasses of tumours. Once this is accomplished, define a procedure that allows classification of new biological specimens.

➲ ***Co-visualize gene expression levels with chromosomal mutation maps***

Combine information from global gene expression analysis and global copy analysis into a map depicting over- and under-expression as well as gene deletions and amplifications in order to explain variations in gene expression.

➲ ***Molecular pathways as meta-analysis descriptors of disease states and subclasses***

Use information about known molecular pathways in order to visualize gene expression data. Identify parameters characteristic of molecular pathways that correlate with clinical behaviour or disease subclass.

5.2 Scenario SC1: A Complex Query Scenario for the TOP Trial

5.2.1 Background

In the **TOP** trial, women with early stage estrogen receptor negative BC receive four cycles of preoperative epirubicin and four cycles of postoperative docetaxel. The biologic hypothesis to be tested is that estrogen receptor negative tumours with topoisomeraseII amplification/overexpression will have a superior response rate to epirubicin. It is hoped from this trial that a molecular signature predicting response or resistance to epirubicin will be able to be established[3] (see figure below).

[3] Sherene Loi, Marc Buyse, Christos Sotiriou and Fatima Cardoso. (2004). Challenges in breast cancer clinical trial design in the postgenomic era. *Current Opinion in Oncology* 16:536–541.

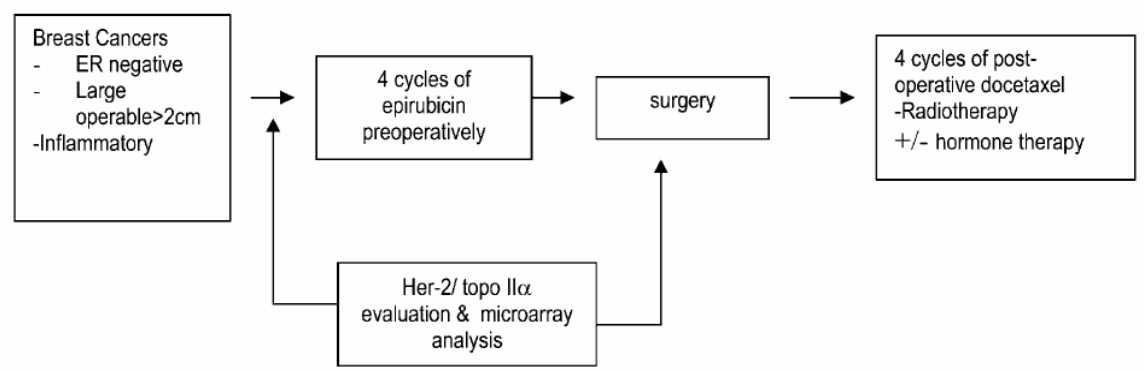


Figure 6: TOP clinico-genomic; hypothesis: topoisomerase IIa_ amplified/overexpressing tumors respond better to anthracyclines (+ identify a molecular signature of anthracycline response/resistance).

5.2.2 Scope and Goals

A researcher involved in the TOP trial tests a hypothesis to explain behavior of non-responders patients who were withdrawn from the trial.

5.2.3 Workflow

In implementing such a scientific analysis a typical user needs to do the following:

1. **Data Access.** Identify the TOP trial patient cases from the UoC (UoC Hospital, Crete) and JBI (Jules Bordet Institute, Belgium) ACGT sites that meet the following clinico-genomic/genetic criteria:

   (i) inflammatory breast cancer that show less than 50% tumor regression, and received less than 1 Epirubicine cycle due to serious adverse event allergy. This implies ***access to and retrieval of data*** from the respective **Clinical Information Systems** in the Cretan and JBI sites; and

   (ii) chromosomal amplification in region 11q, excluding those who show polymorphisms in the specific glucuronidating enzyme of epirubicin UGT2B7. This implies ***access to and retrieval of data*** from the respective **Genetic Information Systems** in the corresponding sites.

2. **Data Access.** Get the pre-operative and post-operative gene expression (microarray) data for the retrieved (from steps **1.**i, ii) patient cases. This implies access to and retrieval of data from the respective *Genomic/Microarray Information Systems* in the Cretan and JBI sites, e.g., BASE-like / MIAME compliant microarray information systems.

   Technology support for steps 1. and 2. *MEDIATION* service (or, set of services) (a) ACGT [breast] *cancer-ontology*; (b) Ontology-based *Query formation*]; (c) Ontology-based/supported *Meta-data* used to describe respective patient-related data sources (i.e., information systems); (d) Clinico-genomic/genetic information systems' *wrappers*; (e) *Query decomposition* (i.e., to retrieve data from multiple, distributed and heterogeneous clinico-genomic/genetic data sources) and formation of the respective

(low-level) *SQL* query structures; and (f) Grid-enabled access to distributed data sources.

3. **Data Pre-processing**

   (i) Identify *common ORFs/Genes* used by the respective microarray specific experiments and filter-out genes that are not in common. This implies utilization of gene converter services based on standard genomic nomenclatures and public data banks, e.g., HUGO, Genbank, Ensembl, etc.

   (ii) *Normalizatio*n of the remaining (after completion of step 2.i) gene-expression data. This implies use of gene expression normalization/transformation tools/services.

4. **Data Analysis**

   Compare pre-operative and post-operative gene-expression data and identify the most discriminatory genes. This implies utilization of *data-mining* tools and services for gene/feature-selection, classification and/or clustering – and respective visualization tools.

   Technology support for steps 3. and 4. *Data analysis/mining* services (a) discover appropriate *Grid-enabled* tools/services for feature-selection, classification, clustering and visualization; (b) select appropriate tools, invoke them and orchestrate them into an analysis workflow.

5. **Genomic Annotation Services**

   (i) Obtain functional annotation for the identified most-discriminatory genes. This implies access to and utilization of *public nomenclatures, ontologies*; and

   (ii) Identify those genes expressed in B-lymphocytes. This implies access to reliable and authenticated *public gene-expression databases*.

   Technology support step 5. (a) Access to *external public genomic databanks* and formation of respective cancer-ontology compliant queries [the respective public data based could be accessed from their original site or, could be downloaded and maintained within the ACGT environment in respective sites]; (b) *Grid-enabled* access and retrieval of public genomic/genetic databases. Algorithms, tools, components for feature-selection, classification and clustering, as well as visualization components; (c) It might be of use to enable (ontology-based) annotation of results.

6. **Identification of Molecular pathways.** Map the identified genes into *regulatory pathways* and find potential *molecular paths* in these pathways. This implies access to public molecular pathways (regulatory and/or metabolic), e.g., KEGG, CyC pathways, etc.

   Technology support step 6. (a) Access to *external molecular pathways; (b) Grid-enabled graph-based, constraint-satisfaction* or other techniques for *reasoning with molecular pathways* (e.g., for the identification of operating / non-operating paths); (c) (ontology/semantic-based) annotation and representation (XML) of results.

7. **Biomedical Literature Search/Mining** Get the literature related to kinases present in specified pathways. This implies discovery, and invocation of appropriate text-mining tools/services.

Technology support step 7. (a) Discovery of appropriate services for access to *biomedical literature information sources; (b) selection of an appropriate one and its invocation* (c) (ontology/semantic-based) annotation and representation of results.

8. **Reporting.** Form and fill-in a standard reporting form for all the performed steps.

Technology support step 7. (a) ACGT cancer and clinical-trials ontology for annotating results and sections of the reports and (b) appropriate end user annotation tools.

5.3 Scenario SC2: Identification of nephroblastoma antigens

5.3.1 Background - abstract from an article of a similar scenario

Reference: Nicole Comtesse, Andrea Zippel, Sascha Walle, Dominik Monz, Christina Backes, Ulrike Fischer, Jens Mayer, Nicole Ludwig, Andreas Hildebrandt, Andreas Keller, Wolf-Ingo Steudel, Hans-Peter Lenhof, Eckart Meese: Complex humoral immune response against a benign tumour: Frequent antibody response against specific antigens as diagnostic targets. PNAS 102:9601-9606, 2005

There are numerous studies on the immune response against malignant human tumours. This study was aimed to address the complexity and specificity of humoral immune response against a benign human tumour. We assembled a panel of 62 meningiomaexpressed antigens that show reactivity with serum antibodies of meningioma patients, including 41 previously uncharacterized antigens by screening of a fetal brain expression library. We tested the panel for reactivity with 48 sera, including sera of patients with common-type, atypical, and anaplastic meningioma, respectively. Meningioma sera detected an average of 14.6 antigens per serum and normal sera an average of 7.8 antigens per serum (P _ 0.0001). We found a decline of seroreactivity with malignancy with a statistical significant difference between common-type and anaplastic meningioma (P<0.05). We detected 17 antigens exclusively with patient sera, including 12 sera that were reactive against KIAA1344, 9 against natural killer tumour recognition (NKTR), and 7 against SRY (sex determining region Y)-box2 (SOX2). More than 80% of meningioma patients had antibodies against at least one of the antigens KIAA1344, SC65, SOX2, and C6orf153. Our results show a highly complex but specific humoral immune response against a benign tumour with a distinct serum reactivity pattern and a decline of complexity with malignancy. The frequent antibody response against specific antigens offers new diagnostic and therapeutic targets for meningioma. We developed a statistical learning method to differentiate sera of meningioma patients from sera of healthy donors.

5.3.2 Scenario description

Wilms tumour is the most common renal tumour in children. Regardless of the efforts in the last 25 years, in the SIOP trials the diagnosis is done by imaging studies alone before starting a preoperative chemotherapy. A definitive diagnosis is available after histological proof after surgery of the tumour. As a disadvantage 1 % of children receive a chemotherapy having benign disease. In this respect, this scenario is highly important for helping to assure the correct diagnosis before starting any kind of treatment.

Immunogenic tumour-associated antigens have been reported for a variety of malignant tumours including brain tumours, prostate, lung and colon cancer. In a first step,

immunogenic Wilms tumour associated antigens will be identified by immunoscreening of a cDNA expression library. Five sera in total from Wilms tumour patients of all three risk groups will be combined and diluted to a final concentration of 1:1000. Antigen-antibody complexes are detected with horseraddish-conjugated anti-human IgG antibody, followed by chemiluminescent detection with ECFTM. This first step will identify those antigens that show reactivity against serum antibodies of patients with Wilms tumour and not with healthy individuals. Only those antigens that react with this pooled serum and not healthy serum (newly identified Wilms tumour antigens); will be used in the following experiments. In step two, serum from a specific patient will be tested against these newly identified Wilms tumour antigens.

As a result in each patient there will be a specific pattern of antigens found, found by the reaction between tumour associated antigen and serum antibody measured by chemiluminescent detection. This specific pattern (different antigens) will be used as a result of the experiment. This pattern will be correlated to the histological subtype of the tumour, the gene expression profiling of the tumour, the response to chemotherapy and the outcome of the patient. As control we will include sera of healthy donors of different age groups and sera of patients with other tumours, like neuroblastoma, that play a role in differential diagnosis.

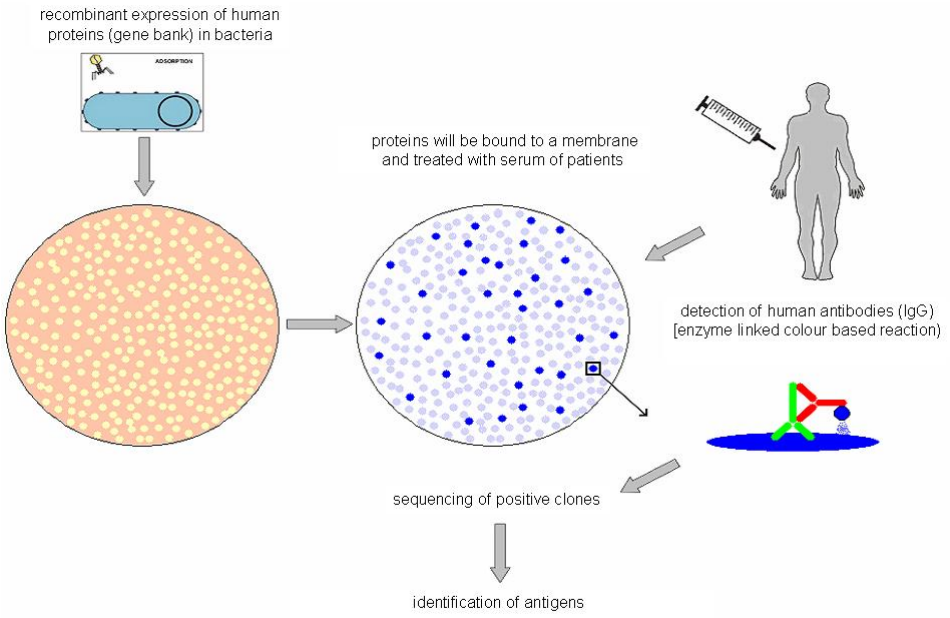


Figure 7: Schematic description of SEREX method

5.3.3 Goals

The pattern of the identified antigens will contribute to answering key questions about the humoral immune response in Wilms tumour patients:

- Are Wilms tumours associated with frequent antibody response?
- Is there a complex and/or specific antibody response?
- Is this response associated with specific genetic features like gene amplifications or DNA losses?

- Do these immunogenic antigens share common features like specific sequence motives?
- Does the seroreactivity pattern allow early identification of Wilms tumours and also their histological subtypes?
- Does the seroreactivity pattern represent a prognostic marker for Wilms tumours in respect to chemotherapeutic response and / or outcome?

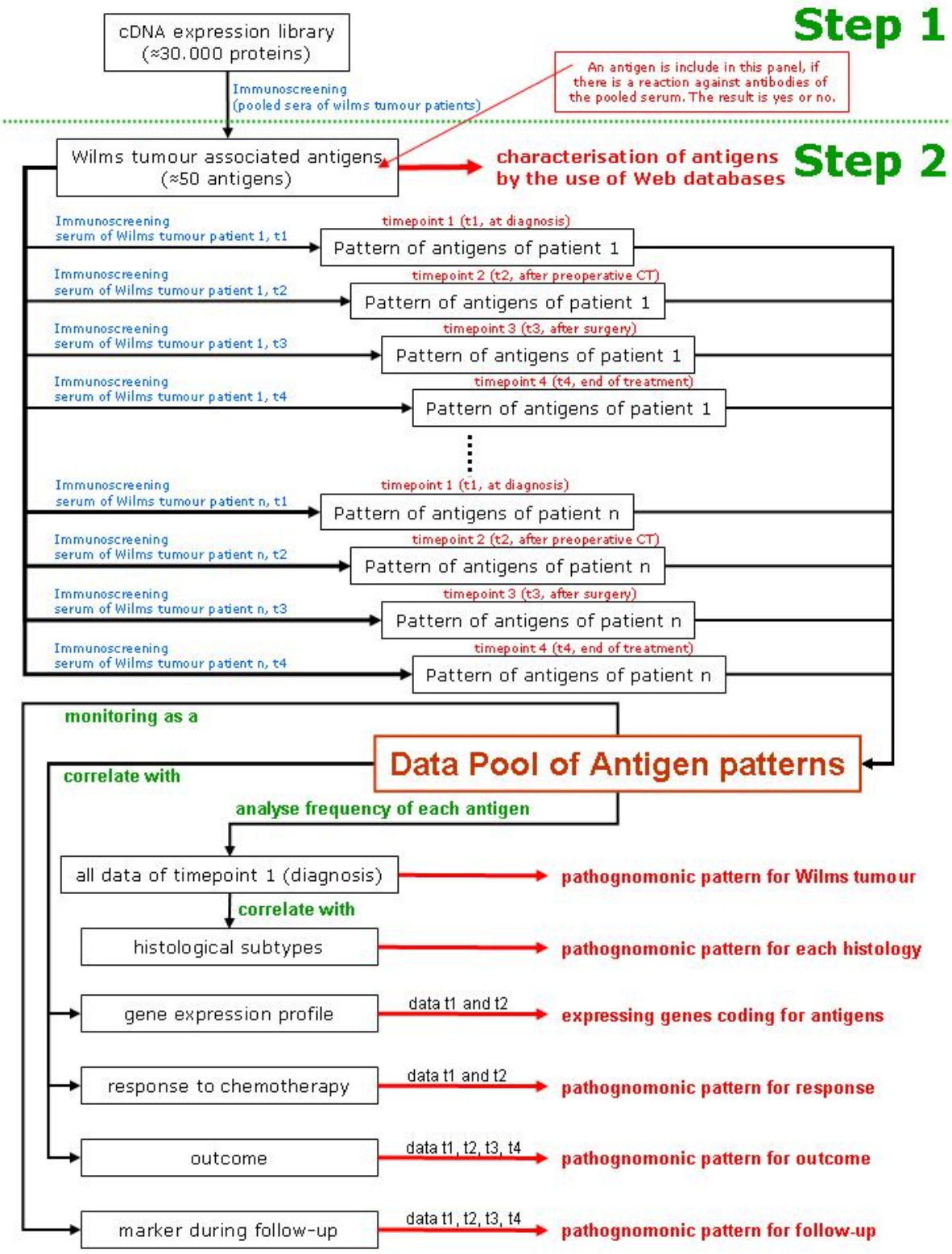


Figure 8: Schematic description of the scenario

5.3.4 Data required

For the implementation of this user scenario several types of data are required. Namely, a Clinical data, proteomic data (SEREX), microarray-data (GMS 417 arrayer [MWG Biotech, Ebersberg, Germany]). Extra data are used from Web-databases.

5.3.5 Description of available data

Clinical data: Clinical database of the SIOP 2001/GPOH

Proteomics data: Excel sheet (see following table)

Data Pool of Antigens	Time point*	antigen 1	antigen 2	antigen 3	antigen 4	antigen 5	…	…	Antigen n
Patient 1	1	+	-	-	+	+	…	…	+
Patient 1	2	+	+	-	-	-	…	…	+
Patient 1	3	+	-	+	+	+	…	…	+
Patient 1	4	+	+	-	-	-	…	…	-
Patient 2	1	-	+	-	+	+	…	…	-
Patient 2	2	+	+	-	-	+	…	…	-
Patient 2	3	-	-	+	+	+	…	…	+
Patient 2	4	+	+	-	+	-	…	…	-
Patient n	1	-	-	+	-	+	…	…	+
Patient n	2	-	+	+	-	-	…	…	+
Patient n	3	+	+	+	+	+	…	…	-
Patient n	4	-	+	-	-	+	…	…	+

* timepoint 1: at diagnosis, without treatment
timepoint 2: after preoperative chemotherapy, before surgery (not available in primarily operated patients)
timepoint 3: after surgery
timepoint 4: at the end of treatment

Microarray data: MIAMExpress (http://www.ebi.ac.uk/miamexpress).

Data available from websites: (see following table)

Information on chromosomal localization, protein function, and sub cellular localization has to be retrieved from	National Center for Biotechnology Information	www.ncbi.nlm.nih.gov
	GeneCards	http://www.genecards.org/index.shtml
	EBI	http://www.ebi.ac.uk/
	Swiss Prot	http://www.ebi.ac.uk/swissprot/access.html
Information on pathways has to be retrieved from	KEGG PATHWAY database	http://www.genome.jp/kegg/pathway.html
Information on pathways has to be retrieved from	Biocarta database	http://www.biocarta.com/search/index.asp
Information on domains has to be retrieved from	SMART database	http://smart.embl-heidelberg.de
Information about antigens found in other tumours from	SEREX database Cancer Immunome database	http://www2.licr.org/CancerImmunomeDB/
	CAP * (Cancer associated	http://www.bioinf.uni-sb.de/CAP/

	proteins) database	
Translation of DNA to Protein	Swiss Institute of Bioinformatics	http://www.expasy.org/tools/dna.html
Information about autoimmunity of antigens from	the autoimmune database	http://www.wiley-vch.de/contents/jc_2040/2005/25481_s.pdf

Relevant literature:

[PIE2004] Pierre Dönnes, Annette Höglund, Marc Sturm, Nicole Comtesse, Christina Backes, Eckart Meese, Oliver Kohlbacher, Hans-Peter Lenhof: Integrative analysis of cancer-related data using CAP. FASEB J 18:1465-1467, 2004

[BAC2005] Backes C, Kuentzer J, Lenhof HP, Comtesse N, Meese E: GraBCas: a bioinformatics tool for score-based prediction of Caspase- and Granzyme B-cleavage sites in protein sequences. Nucleic Acids Research 33: W208-W213, 2005

5.3.6 Workflow

The forseen sequence of activities for the execution of the scenario are as follows:

5.3.6.1 Step 1

- Clinical data are not needed for step1.
- Data of the SEREX experiments will be sent to the ACGT platform in form of an Excel sheet. In Step 1 only the ID numbers and the nucleotide sequences of positive clones are included.
- Nucleotide sequences will be given to the translation tool of Expasy (http://www.expasy.org/tools/dna.html) and translated into six possible reading frames (3 reading frames from 3`to 5` and 3 frames from 5`to 3`).
- The used frame will be found by the vector amino acid sequence. This is the first protein sequence according to the clone in the experiment. This protein will be used later again.
- The nucleotide sequence of the positive clone that was found in the experiment will be analysed with the NCBI web tool BLAST (http://www.ncbi.nlm.nih.gov/BLAST/ - following nucleotide to nucleotide (blastn)) at this page either. BLAST will start by clicking Blast button. A request ID will be given and search will continue by clicking “Format”
- The most similar sequences will be received by this search. Human genes (NCBI accession number in format NM_xxxx or XM_xxxx) have to be selected. The most identical sequence will be given at the top of the search results.
- By choosing the most similar gene the information about gene name (given in the definition section) and gene symbol (in brackets in the section gene). The protein ID (NP_XXXX at NCBI) is given at the bottom of this page and is directly linked to the correlating NCBI page. It is also possible to get links to diseases and genetic disorders

linked to these genes in the Online Mendelian Inheritance in Man™ database on NCBI by choosing the MIM link at the point gene.

- The linked protein page has to be selected and on the connected page the information about the expressed protein will be given.
- To compare the protein sequence corresponding to the gene and the protein sequence corresponding to the nucleotide sequence of the identified clone the protein sequence found in NCBI has to be in FASTA format. This will be done automatically by selecting FASTA in the task line at “Display" on the page the protein was found in NCBI.
- The two protein sequences are aligned at the NCBI → Blast→Special→blast2seq page (http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi)
- The sequence identified with the translation tool at Expasy (1st) is given to the field of sequence 1, the protein sequence found in NCBI (2nd) is given to the field of sequence 2.
- Because two proteins are compared, the program task line at the top of the page has to be changed from blastn (for nucleotides) to blastp (proteins).
- The comparison is necessary to ensure that the protein that is expressed by the clone is at least partly corresponding to the expressed protein expected from the nucleotide sequence homology.
- When the sequence found in the expasy tool shows significant similarity to the really expressed and expected protein it is called in frame. Nevertheless the clones with unknown frames may express existing proteins either. Homologous proteins can be searched via further database analysis (e.g. at the NCBI database either).
- There may be truly expressed proteins homologous to protein expressed by the not in frame clones but it is necessary to proceed further with analyses to get a conclusion about their function. This will be probably done in a future scenario.
- Further characterizations of the genes and the proteins are possible by the use of the databases described in the table above. The NCBI accession number is used as identification. The NCBI database also provides links to publications of each of the antigens.
    - Antigen function, localisation in the cell and the intracellular process is characterized by using the entrez gene from the NCBI homepage. (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=search&term=)
    - Correlating pathways of the antigen is provided by the KEGG database using the gene symbol or the gene ID received before at NCBI.
    - Similar to the KEGG database a search has to be done at Biocarta.
    - The SMART database is used to show the protein domains
    - The SEREX and the CAP database are analysed to detect other tumours expressing the clones and proteins found in the actual experiment.
    - The relation between the expressed genes and the chromosomal localisation as well as the expression in different human tissues is found in the gene cards.
    - Detailed information from other homepages concerning the isolated gene at gene cards too is seeked.

- Further information concerning chromosomal localization, protein function, and sub cellular localization is retrieved from EBI and SwissProt.
- Users may receive an accession number of EBI by entering the gene symbol in the overall search of the SwissProt homepage (http://www.ebi.ac.uk/swissprot/access.html).

- The results of the different databases will be summarized and visualized in a simple and clear way. A filter-tool including search criteria will help to present the results in a flexible way.

5.3.6.2 Step 2

In the second step the results of the individual experiments described before must be analysed, correlated and visualized in the following way. The number of used antigens is defined by the result of step 1 (only positive clones with the pooled serum).

- Clinical data is collected by the clinicians and pseudonymized / anonymized. The database of the SIOP 2001/ GPOH is used.

- The data of the antigen/antibody experiment is provided to the ACGT platform in the form of an excel sheet as shown in the table of proteomics data.

- Positive antigene-antibody reactions have to be described according to descriptive statistical tools (tables, bars, box-plots, etc) and a statistical test battery (t-Test, etc). Statistical tests are used to find significant differences between different groupings of the data according to clinical questions. Some of the analysis are listed below:
  - the percentage of positivity of each clone in summary and
  - for the same histological subtype and
  - in comparison of different histological subtypes
  - at the 4 different time points and
  - in comparison to the different time points
  - in correlation to molecular biological findings (gene signature)
  - to the clinical outcome
- Positivity of clonesare described as a function over time in individual patients and in clinically defined groups of patients, to answer the question if the antigen/antibody pattern of positivity can be used as a tumour marker.

5.3.7 Stakeholders Profile

The users are IT literate, experienced clinical researchers, namely:

Prof. Dr. Eckhart Meese, Department of Human Genetics and Molecular Biology, UdS
Prof. Dr. Manfred Gessler, Physiologische Chemie I, Universität Würzburg
Prof. Dr. Norbert Graf, Alexander Hoppe, UdS

5.4 Scenario SC3: Correlating phenotypical and genotypical profiles

The presented scenario could be extended to the ACGT TOP breast-cancer trial.

5.4.1 Background

"… Two major challenges in using genomics for disease (such as breast cancer) diagnostics are the ability to find robust classifications that maintain prognostic significance across different patient populations, and the ability to *effectively translate those classifications into the clinical laboratory* **…"**[4].

" … Although *it is still of great importance to gain an understanding of which patterns of gene expression are linked to known variables* such as ER-α status, studies should be designed to reveal rather than obscure these links, and *to uncover any potential gene expression patterns that predict outcome within uniform groups* …"[5]

5.4.2 Scenario description

This is an Integrated Clinico-Genomics Knowledge Discovery Scenario. Its realization will be based on a multi-strategy data-mining process.

The fundamental objective is to offer a flexible and effective data analysis framework – based on the smooth integration of interoperable, high-performing and discoverable (i.e., Gridfied) data mining operations and corresponding tools.

The aim is: (a) to support the semantic integration of a patients' clinical and genomic data, and (b) to reveal interesting relations between the relevant data sources towards the composition of interesting and individualized clinico-genomic patients' profiles.

5.4.3 Required and available data

Initial realization (*feasibility* study) of the scenario will be based on relevant *public-domain* data. In this context we will focus on the ***NKI*** dataset which records both clinical, histopathology, treatment and clinical outcome as well as respective gene-expression data for 295 **breast-cancer** patient cases. All relevant data may be downloaded from http://microarray-pubs.stanford.edu/wound_NKI/explore.html.

Sebsequently, when the ACGT pilot sites have begun to collect the appropriate datasets, the scenario will be realized based on these actual datasets.

5.4.4 Workflow

The forseen sequence of activities for the execution of the scenario are as follows:

[4] Laurent Perreard, et al. Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. *Breast Cancer Research* 2006, **8:**R23, doi:10.1186/bcr1399, http://breast-cancer-research.com/content/8/2/R23.

[5] Sofia K Gruvberger, et al. Expression profiling to predict outcome in breast cancer: the influence of sample selection. *Breast Cancer Res* 2003, 5:23-26, doi:10.1186/bcr548, http://breast-cancer-research.com/content/5/1/023.

Step-1. The clinical, histopathology and the gene-expression data of all NKI patients are appropriately entered into the relevant clinical and gene-expression information systems (see below "Technical Requirements")[6].

Step-2. The information systems are queried to retrieve (query-speific) patients' clinico-histopathology and gene-expression data from the relevant information system sources (see below "Technical Requirements"). The results are persistently stored.

Step-3. The samples may be assigned, by the involved clinical users, to various clinico-histopathological categories corresponding to specific profiles, e.g., profiles or classes referring to specific tumour types, stages, drug response statuses etc.

Besides the investigator-guided process, advanced data-mining operations are utilised in order to automatically discover indicative CHPPs. Different data-mining operations are used, such as (i) ***Descriptive CHPPs*****.** Clustering of patients/samples to categories of 'similar clinico-histopathology profiles'. This helps in the identification of potentially interesting cohorts to target a clinico-genomic study. Among others, the clustering operations should support feature focusing in assisting the user to focus on specific clinio-histopathological features of interest. (ii) ***Discriminant CHPPs***. When a decision-feature is selected by the user (e.g., "metastasis vs. no-metastasis patients") then, with the aid of classification and feature-selection techniques and algorithms the user is assisted in identifying combinations of clinico-histopathological feature-value combinations being able to discriminate between pre-specified patient groups, e.g., *"good/bad prognostic clinico-histopathological profiles"*. These CHPPs are contrasted with respective gene-expression profiles and assess correspondences and respective diagnostic/prognostic prediction performances (see step 4 below).

Step-4. By measuring transcription (gene expression) levels of genes in an organism under various conditions and in different tissues, Gene Expression Phenotypical Profiles are constructed or, patterns - GEPP, which characterize the dynamic functioning of each gene in the genome are identified. The identification of patterns in complex gene expression datasets provides two benefits: (a) generation of insight into gene transcription conditions; and (b) characterization of multiple gene expression profiles in complex biological processes, e.g. pathological states.

Step-5. The next step now is to link and relate the two phenotypic profiling types. For achieving this the smooth integration of data-mining operations, i.e, clustering, association rules mining and feature-selection with classification operations, is required. We refer to this process as I-CGKD: Integrated Clinico-Genomic Knowledge Discovery. The process unfolds as follows:

> At first a user utilises an appropriate (unsupervised) clustering method in order to identify clusters of genes or, metagenes, based on their gene-expression profiles. The objective is mainly to reduce the dimensionality of the search space (i.e., from 1000ts of genes to 10ths of metagenes). These metagenes represent potentially interesting GEPPs. The following question then is: "does these GEPPs relate to some potential CHPPs, and if yse how?". In responding to this question the user invoces appropriate tools, i.s. an association rules mining (ARM) tool in order to automatically discover 'highly confident' correlations between GEPPs and CHPPs. Such correlations assist in a potential re-classification of the targeted disease in the sense

[6] In the actual version (as to be supported by the integrated ACGT final architecture and respective services) this step will be abandoned based on the availability of patient data in the respective clinical and microarray information systems.

that specific (strong) GEPPs point to specific CHPPs and respective behaviours. Moreover, if the description of the linked and identified CHPPs include a 'decision variable' then, the user initiates a gene-selection process on the population of samples being covered by the CHPPs and on the genes included in the GEPPs (i.e., genes in metagenes). In other words, we are now seeking for 'individualised' molecular/genomic signatures.

5.4.5 Technical Requirements

Realization of steps 1 and 2 of the presented scenario and the I-CGKD process is depended on the availability of appropriate information systems to store and maintain respective patients' clinical and gene-expression data as well as to a mediation infrastructure to retrieve data from (potential) distributed and heterogeneous data sources.

- ***Clinical Information Systems.*** All relevant patients' data are stored, managed and retrieved in appropriate *Clinical Information Systems* (CIS)..

- ***Gene-Expression Information Systems.*** A specific microarray-experimentation and gene-expression profiling information system is utilised for the management of the gene expression data. The system is compliant to the MIAME – Minimum Information About Microarray Experiments which represents a de-facto standard for microarray experiments and gene-expression profiling data (http://www.mged.org/Workgroups/MIAME/miame.html).

- ***Data Mediation Services.*** A middleware layer for *seamless* access and integration of data is required, supporting the biomedical investigator to form *combined clinico-genomic queries* (e.g., *" get the gene-expression profiles for 'microarray experiments meeting specific criteria' of all breast cancer diseased women meeting a clinical profile of: 'age over 40', 'ER+', 'lymph-node –' and 'followd-up for over 5 years' "*.

  Query results are combined and a standard data-enriched XML file is created for further utilization, e.g., for data-mining/ knowledge-discovery operations.

Realization of the remaining steps, i.e. the I-CGKD process, is depended on the availability of the appropriate data mining operations and appropriate visualization tools.

The proposed process is inspired by respective multi-strategy machine learning and case-based reasoning methodologies. The whole approach, and the actual realization of the proposed scenario, is based on the smooth integration of three distinct data-mining analytical tools. Namely:

- ***Clustering tool.*** A k-means clustering algorithm operating on categorical data is required. With this approach the clusters of genes that best describes the available patient cases are selected, i.e., clusters that cover an adequate number of genes and for which an adequate number of samples shows significant 'low' or, 'high' gene-expression values; we refer to them as metagenes.

- ***Association Rules Mining tool.*** Aiming at discovering of 'causal' relations (rules with high confidence) between genes (actual clusters of genes, i.e., the metagenes) and patients' phenotypic profiles (i.e., between CHPPs and GEPPs).

- ***Feature Selection and Classification tool.*** For the selection of the most discriminant genes, i.e., genes being able to distinguish (i.e., with high accuracy)

between patients' pre-specified classes, i.e., decision variables of the focused patient cases (disease state, survival category, etc).

5.4.6 References

[KAN2006] Kanterakis A. and Potamias G. Supporting Clinico-Genomic Knowledge Discovery: A Multi-strategy Data Mining Process. In G. Antoniou, et al. (Eds.): SETN 2006, LNAI 3955, pp. 520-524, 2006. (http://www.springerlink.com/(mibwxn45mifi5r45aa5ddp2q)/app/home/content.asp?referrer=contribution&format=2&page=1&pagecount=0).

[POT2005] Potamias G., et al. Breast Cancer and Biomedical Informatics: The PrognoChip Project. In Proceedings of the 17th IMACS World Congress Scientific Computation, Applied Mathematics and Simulation, Paris, France, 2005. (http://sab.sscc.ru/imacs2005/papers/T3-I-68-1066.pdf)

[POT2004a] Potamias G, Koumakis L., and Moustakis V. Gene Selection via Discretized Gene Expression Profiles and Greedy Feature-Elimination. In G.A. Vouros and T. Panayiotopoulos (Eds.): SETN 2004, LNAI 3025, pp. 256–266, 2004. (http://www.springerlink.com/(mibwxn45mifi5r45aa5ddp2q)/app/home/content.asp?referrer=contribution&format=2&page=1&pagecount=11).

[POT2004b] Potamias, G., Koumakis, L. (2004). HealthObs: An Integrated System for Mining Clinical XML-formatted Data. In 2nd International Conference on Information Communication Technologies in Health, Journal for Quality of Life Research (JQLR) (http://www.ics.forth.gr/~potamias/WWW_Potamias_ResearcherB/PUBS/C.16.zip).

5.5 Scenario SC4: Reporting of Adverse Events and Severe Adverse Reactions

Pharmacovigilance is defined as the pharmacological science relating to the detection, assessment, understanding and prevention of adverse effects, particularly long term and short term side effect, of medicines. It is gaining importance for doctors and scientists as the number of stories in the media of drug recalls increases. Because clinical trials involve, at most, several thousand patients, less common side effects and Adverse Drug Reactions (ADRs) are often unknown at the time a drug enters the market. Even very severe ADRs, such as liver damage, are often undetected because study populations are small. Postmarketing pharmacovigilance uses tools such as data mining and investigation of case reports to identfy the relationships between drugs and ADRs.

The pharmacovigilance effort In Europe is coordinated by the European Medicines Agency (EMEA) and conducted by the national competent medicines authorities (NCA). The main responsibility of the EMEA is to maintain and develop the pharmacovigilance database consisting of all suspected serious adverse reactions to medicine observed in the European community. The system is called Eudravigilance [http://www.eudravigilance.org/human/index.asp] and contains separate but similar databases of human and veterinary reactions.

All clinical trials within the European Community have to be done according to the regulations of the Directive 2001/20/EC. By 20 November 2005 new European legislation requires to submit all received adverse reactions in electronic form. This can be done with commercial software developed for the purpose or with a web utility called EVWEB accessible through the EudraVigilance homepage. Registration for use of EVWEB is nessesary.

5.5.1 Background

The responsibilities of the investigator in relation to the notification of Adverse Events (AEs) are set out in this Directive: “The investigator shall report all Serious Adverse Events (SAEs) immediately to the sponsor except for those that the protocol or investigator's brochure identifies as not requiring immediate reporting. The initial report shall be promptly followed by detailed, written reports. The initial and follow-up reports shall identify the trial subjects by unique code numbers assigned to the latter.”

Adverse events and/or laboratory abnormalities identified in the protocol as critical to the evaluation of safety must be reported to the sponsor by the investigator according to the reporting requirements within the time periods specified in the protocol. The investigator shall supply the sponsor and the Ethics Committee with any additional requested information, notably for reported deaths of a subject.

The sponsor of a clinical trial has the obligation to report all Severe Adverse Events (SAEs) and Suspected Unexpected Severe Adverse Reactions (SUSARs) to the legal authorities, the ethical committees and the participating centres as stated in the directive:” The sponsor is responsible for the prompt notification to all concerned investigator(s), the Ethics Committee and competent authority of each concerned Member State of findings that could adversely affect the health of subjects, impact on the conduct of the trial or alter the competent authority’s authorisation to continue the trial in accordance with Directive 2001/20/EC“.

A detailed guidance on the collection, verification and presentation of adverse event/reaction reports, together with decoding procedures for unexpected serious adverse reactions is published:
http://eudract.emea.eu.int/docs/Detailed%20guidance%20collection%20of%20adverse%20events.pdf

5.5.2 Goals

Stakeholder benefit	*Supporting features*
Easier work for Investigators (clinicians)	only one reporting system, independent from the trial
Less work for Sponsor and legal authorities	They will get reports in a standardized way via online access
Improved Patient security	Reports are standardized and in the same manner independent from the trial

5.5.3 Required Datasets and Tools

For the implementation of this scenario the following datasets and tools are required:

- EudraCT Database
- MedDRA Database
  - The latest version should be applied, using version 4.1 or later versions. Lower level terms (LLT) should be used)
- Clinical Database
- A tool that can easily accessed and used by the stakeholders

5.5.4 Description of Available data and EudraVigilance

The data of the EudraCT and MedDRA databases are standardized. The data from the clinical database are depending on the clinical trials database. The needed data are described in section 1.1.8.

Further information can be obtained from the EudraVigilance homepge [http://www.eudravigilance.org/human/index.asp]. EudraVigilance is a data processing network and management system for reporting and evaluating suspected adverse reactions during the development and following the marketing authorisation of medicinal products in the European Economic Area (EEA). The first operating version was launched in December 2001.

EudraVigilance supports in particular the:

- Electronic exchange of suspected adverse reaction reports (referred to as Individual Case Safety Reports) between the European Medicines Agency (EMEA), national Competent Authorities, marketing authorisation holders, and sponsors of clinical trials in the EEA;
- Early detection of possible safety signals associated with medicinal products for Human Use;
- Continuous monitoring and evaluation of potential safety issues in relation to reported adverse reactions;
- Decision making process, based on a broader knowledge of the adverse reaction profile of medicinal products especially in the frame of Risk Management.

Taking into account the pharmacovigilance activities in the pre- and post- authorisation phase, EudraVigilance provides two reporting modules:

- **The EudraVigilance Clinical Trial Module (EVCTM)** to facilitate the electronic reporting of Suspected Unexpected Serious Adverse Reactions (SUSARs) as required by Directive 2001/20/EC.

The EudraVigilance Post-Authorisation Module (EVPM) designed for post-authorisation ICSRs, Regulation (EC) No 726/2004, Directive 2001/83/EC as amended, and Volume 9 of the "Rules Governing Medicinal Products in the European Union".

5.5.5 Workflow

The following steps present the process for the implementation of the scenario and reveal the different services the platform has to provide in order to support such scenario.

1. Adverse events (AEs) / Severe Adverse Reaction / Suspected Unsuspected Severe Adverse Reactions (SUSAR`s) and/or laboratory abnormalities identified in the protocol as critical to safety evaluations are reported to the sponsor according to the reporting requirements and within the time periods specified in the protocol. They have to be send by the investigator (participating center) to the sponsor by using the ACGT platform in a standardized format.

2. Inform the sponsor about the entry of a new event by the ACGT platform automatically and send the report to the sponsor by email.

3. The sponsor has to evaluate every AE regarding
   - its seriousness and
   - the causality between the investigational medicinal product(s) and/or concomitant therapy and the adverse event
4. After evaluation of the AE the sponsor himself confirms the data of the report using the ACGT platform.
5. If the event is confirmed by the sponsor, the report has to be send to all participating centres, the competent authorities and the Ethics Committee.

6. In addition, it is recommended that the sponsor reports them to the marketing authorisation holder and sends the previous notification to the competent authority.
7. The sponsor has to keep detailed records of all AEs reported to him by the investigator(s´).
8. On request of a competent authority in whose territory the clinical trial is being conducted, the sponsor submits detailed records of all adverse events which are reported to him by the relevant investigator(s´).
9. All the events reported (with special emphasis of SUSAR`s) are presented as a line listing or in a table for the annual report.

10. The annual report is sent to the participating centers, Ethical Committee, Legal authorities.

11. It is of utmost importance that identifiers are used for the trial (EudraCT No.) and the subject (patient). Double reports will be avoided by these identifiers and the data given to the database.

12. The reporting of SAEs and SUSARs has to be done in a short time frame defined in the European Directive 2001/20/EC:

    ➲ Fatal or life-threatening SUSARs

    - The competent authority and the Ethics Committee in the concerned Member States should be notified as soon as possible but no later than 7 calendar days after the sponsor has first knowledge of the minimum criteria for expedited reporting. In each case relevant follow-up information should be sought and a report completed as soon as possible. It should be communicated to the competent authority and the Ethics Committee in the concerned Member States within an additional eight calendar days.

    ➲ Non fatal and non life-threatening SUSARs

    - All non fatal and non life-threatening SUSARs and safety issues must be reported to the competent authority and the Ethics Committee in the

concerned Member States as soon as possible but no later than 15 calendar days after the sponsor has first knowledge of the minimum criteria for expedited reporting. Further relevant follow-up information should be given as soon as possible.

13. For reporting death of a subject, the investigator supplies the sponsor and the Ethics Committee with any additional information requested.

5.5.6 Data protection requirements

The Community standards of confidentiality must always be maintained and any relevant national legislation on data protection must be followed.

5.5.7 Format of the standardized SUSAR report

Electronic reporting should be the expected method for expedited reporting of SUSARs to the competent authority. The format and content as defined by the Guidance (http://eudract.emea.eu.int/docs/Detailed%20guidance%20SUSAR.pdf) should be adhered to. The CIOMS-I form is a widely accepted standard for expedited adverse reactions reporting. The standardized report has to be developed for availability on the ACGT platform. The following parameters are needed:

5.5.7.1 Minimum criteria for initial expedited reporting of SUSARs

Information on the final description and evaluation of an adverse reaction report may not be available within the required time frames for reporting. For regulatory purposes, initial expedited reports should be submitted within the time limits as soon as the minimum following criteria are met:

1. a suspected investigational medicinal product,
2. an identifiable subject (e.g. study subject code number),
3. an adverse event assessed as serious and unexpected, and for which there is a reasonable suspected causal relationship,
4. an identifiable reporting source, and, when available and applicable:
    - o an unique clinical trial identification (EudraCT number)
    - o an unique case identification (i.e. sponsor's case identification number).

5.5.7.2 Follow-up reports of SUSARs

In case of incomplete information at the time of initial reporting, all the appropriate information for an adequate analysis of causality should be actively sought from the reporter or other available sources. The sponsor should report further relevant information after receipt as follow-up reports. In certain cases, it may be appropriate to conduct follow-up of the long-term outcome of a particular reaction.

5.5.7.3 Data Elements for SUSAR report

1. Clinical trial identification:

- Clinical trial identification is done by the EudraCT number

2. Subject's details :

- Sponsor's subject identification number,
- Initials, if applicable,

- Gender,
- Age and/or date of birth,
- Weight,
- Height,

3. Suspected investigational medicinal product(s) (IMPs) :

- Name of the IMP or brand name as reported,
- International non-proprietary name (INN),
- Batch number,
- Indication(s) for which suspect investigational medicinal product was prescribed or tested,
- Dosage form and strength,
- Daily dose and regimen (specify units e.g. mg, ml, mg/kg),
- Route of administration,
- Starting date and time of day,
- Stopping date and time, or duration of treatment
    - Unblinding : yes/no/not applicable ; results
    - Investigator's causality assessment
    - Sponsor's causality assessment
- Comments, if relevant
    - (e.g. causality assessment if the sponsor disagrees with the reporter; concomitant medications suspected to play a role in the reactions directly or by interaction; indication treated with suspect drug(s)).

4. Other treatment(s) :

- For concomitant medicinal products (including non prescription/OTC medicinal products) and non-medicinal product therapies, provide the same information as listed above for the suspected investigational medicinal product.

5. Details of suspected Adverse Drug Reaction(s) (ADRs) :

- Full description of reaction (s) including body site and severity, as well as the criterion (or criteria) for regarding the report as serious should be given. In addition to a description of the reported signs and symptoms, whenever possible attempts should be made to establish a specific diagnosis for the reaction.
- Reaction(s) in MedDRA terminology (lowest level term)
- Start date (and time) of onset of the reaction,
- Stop date (and time) or duration of the reaction,
- De-challenge and re-challenge information,
- Setting (e.g. hospital, out-patient clinic, home, nursing home),
- Outcome : information on recovery and any sequelae; what specific tests and/or treatment may have been required and their results ; for a fatal outcome, cause of death and a comment on its possible relationship to the suspected reaction should be provided. Any autopsy or other post-mortem findings (including a coroner's report) should also be provided when available.
- Other information : anything relevant to facilitate assessment of the case, such as medical history including allergy, drug or alcohol abuse; family history; findings from special investigations.

6. Details on reporter of event/suspected ADR :

- name,
- address,

- telephone number,
- profession (speciality)

7. Administrative and Sponsor details:

- Date of this report
- Source of report: from a clinical trial (provide details if not in Eudract, from the literature (provide copy), spontaneous, other)
- Date event report was first received by sponsor,
- Country in which reaction occurred,
- Type of report filed to authorities : initial or follow-up (first, second, etc),
- Name and address of sponsor/manufacturer/company,
- Name, address, telephone number and fax number of contact person in reporting sponsor,
- identifying regulatory code or number for marketing authorisation dossier or clinical investigation process for the suspected product (for example IND number, NDA number)
- Case reference number (sponsor's/manufacturer's identification number for the case) (this number must be the same for the initial and follow-up reports on the same case).
- Coding, that the sponsor has reviewed the SAE or SUSAR

5.5.7.4 Content of line listing

The line listing identifiable by the sponsor listing reference number or date and time of printing should include the following information per case

- clinical trial identification,
- Study subjects identification number in the trial
- case reference number (Case-ID-Number) in the sponsor's safety database for medicinal products
- country in which case occurred
- age and sex of trial subject
- daily dose of investigational medicinal product, (and, when relevant, dosage form and route of administration)
- date of onset of the adverse reaction.
- If not available, best estimate of time to onset from therapy initiation. For an ADR known to occur after cessation of therapy, estimate of time lag if possible.
- dates of treatment. (if not available, best estimate of treatment duration.)
- adverse reaction : description of reaction as reported, and when necessary as interpreted
- by the sponsor ; where medically appropriate, signs and symptoms can be lumped into diagnoses. MedDRA should be used.
- patient's outcome (e.g. resolved, fatal, improved, sequelae, unknown). This field should indicate the consequences of the reaction(s) for the patient, using the worst of the different outcomes for multiple reactions
- comments, if relevant
    - (e.g. causality assessment if the sponsor disagrees with the reporter; concomitant medications suspected to play a role in the reactions directly or by interaction; indication treated with suspect drug(s); dechallenge / rechallenge results if available)
    - unblinding results in the case of unblinded SUSARs expectedness at the time of the occurrence of the suspected SARs, assessed with the

reference document (i.e. investigator's brochure) in force at the beginning of the period covered by the report.

5.5.7.5 Table for the annual reporting to the Ethical Committee and the legal authorities:

Number of reports by terms (signs, symptoms and diagnoses) for the trial n° : (An * indicates an example of a SUSAR)

Body system / ADR term	Verum	Placebo	Blinded
CNS			
Hallucinations *	2	2	0
Confusion *	1	1	0
Sub-total	3	3	0
CV			
…			
Sub-total			

5.5.8 Stakeholder's Profile

The following stakeholders are involved in this service:

- (Principal) investigator
- Sponsor
- Participating center in the trial
- Ethical committees
- Legal authorities
- Eudract Database

The profile of the each stakeholder and their requirements from the system are shown below.

User	Responsibility	Success criteria	Deliverables
Principal Investigator	To send AEs and/or SARs to the sponsor	Timely receipt of the standardized report by the sponsor	Standardized report
Sponsor	To send the received reports of the principal investigator to the participating centers, the ethical committee and the legal authorities To send all other reports as described	Timely receipt of the standardized report by participating centers, the ethical committees, the legal authorities and in the EudraCT Database	Standardized report

	above, including the annual report		
Participating centers in the trial (Investigators)	To send all SAEs and SUSARs to the sponsor To receive all reports as described above	Timely receipt by the sponsor Receipt of the reports	Standardized report
Ethical Committee	To receive all reports as described above from the sponsor	Receipt of the reports	
Legal authorities	To receive all reports as described above from the sponsor	Receipt of the reports	
EudraCT Database http://eudract.emea.eu.int/	To enter the data of the reports directly into the database	Receipt of the data	

5.6 Scenario SC5: In-silico modelling of tumor response to therapy

5.6.1 Background

The advancement and clinical validation, adaptation and utilization of *in* ***silico*** **(computational) oncology** is an important domain in ACGT project. The aim is to provide clinicians with a decision support tool able to simulate within defined reliability limits the response of a solid tumour to therapeutic interventions based on the individual patient's data. The treatment effects on the normal tissues will also be taken into account even in considerably less detail. An *intermediate goal* of this action is to provide researchers with a versatile platform for integrating experimental and clinical knowledge and performing exploratory experiments *in silico* (on the computer). Therefore, the proposed system is expected to become a prototype *multi-level* cancer biology *integrator*.

Although extensive exploitation of relevant previous work done by ACGT members will take place, large scale extensions and modifications will be implemented in order to cope with the particularly high demands and intricacies of the two clinical cases addressed by ACGT i.e. nephroblastoma (Wilm's tumor) and breast cancer. To this end a computational system denoted by the specially coined term **"Oncosimulator"** will be developed.

As the clinical validation of the "Oncosimulator" will be based on the two clinical studies incorporated in ACGT (nephroblastoma SIOP 2001/GPOH and breast cancer TOP study), the term "*In Silico* Oncology trial" which is sometimes used in the ACGT context actually refers to a *"metatrial"* i.e. a validation procedure aiming at checking and optimizing a complex simulation system through the *observation* of the *time course* of the corresponding physical system's behaviour (here the tumour).

To the best of our knowledge, up to now there have not been any *specially planned*, *large scale*, *molecular biology enhanced clinical trials* [or more correctly clinical validation procedures] in order to test and adapt mathematical or computational models of tumour response to therapeutic modalities.

5.6.2 Goals

The **objective of the study** is to validate, clinically adapt and optimize the "Oncosimulator" for the special cases of nephroblastoma and breast cancer.

To this end:

- for the case of nephroblastoma - the clinical, imaging and molecular data of the patient or
- for the case of breast cancer - the clinical, imaging, histopathologic and molecular data of the patient

following preprocessing will be introduced into the "Oncosimulator" along with the description of the therapeutic scheme (temporal drug administration scheme) to be simulated. The prediction of the "Oncosimulator" regarding the tumour response as a function of time will be compared with the imaging data at various instants during and after the chemotherapeutic scheme. The outcome of the comparison will be used as an adaptation / optimization feedback for the "Oncosimulator".

5.6.3 Data required for the nephroblastoma study

Implementattion of this scenario requires access to the following types of data.

5.6.3.1 Clinical and Molecular Data

- **Clinical Data**
  - Age
  - Sex
  - Weight
  - Height
  - Syndromes (WAGR, Denys-Drash, Beckwith-Wiedermann)
  - Family history
  - Blood cell counts (BCC) {to monitor adverse effects on normal tissues}
- **Imaging Data** (baseline: just before chemotherapy start)
  - CT (DICOM) and/or MRI (DICOM) and/or ultrasound (DICOM)]
  - Three ellipsoidal axes of the tumour.
  - Delineation of the necrotic, cystic, hemorrhagic and solid tumour regions on the tomograhic slices.

- **Molecular Data**
  - Profiling of antibodies to tumour antigens (antigen scenario)
  - Estimated cell type composition of the tumour
  - Estimated tumour cell responsiveness to the drugs under consideration

5.6.3.2 Process and Recommended Treatment Scheme(s) Data

- The above mentioned data is accessed and entered into the "Oncosimulator" which performs the tumour response to chemotherapy simulation.
- A rough estimation of the response of representative normal tissues is also made.
- As a result, the most probable outcome is predicted.
- Based on the "Oncosimulator" prediction (mainly the expected tumour shrinkage), the clinician judges whether or not the chemotherapy outcome would be beneficial to the patient under consideration by also taking into account his or her logic, expertise and even intuition.
- In case that the expected outcome is not judged as beneficial, the patient may proceed to surgery. Otherwise, the chemotherapeutic scheme is applied on the real patient.
- The actual chemotherapy administration schedule is registered.

The following examinations are carried out during and after treatment:

- **During chemotherapy**
  - Ultrasound imaging every week,
  - Recording of the 3 tumour ellipsoidal axes
- **After completion of chemotherapy**
  - Profiling of serum antibodies against tumour antigens
  - CT (DICOM) and/or MRI (DICOM) and/or ultrasound (DICOM)
  - Three ellipsoidal axes of the tumour.
  - Delineation of the necrotic, cystic, hemorrhagic and solid tumour regions on the tomograhic slices.
  - Blood Cell Counts (BCC)
- **After surgery**
  - Histology (types)

    The predicted and the actual **outcome** and **histology** are compared and if they are in significant contradiction an optimization and adaptation loop for the "Oncosimulator" is carried out, otherwise the current checking of the "Oncosimulator" is judged as favourable.

5.6.4 Data required for the TOP breast cancer study

5.6.4.1 Clinical and Molecular Data

For the development and validation of the "Oncpsimulator" for the TOP study, the following types of data are reuired.

- **Clinical Data**
  - Age
  - Sex
  - Weight
  - Height
  - Previous treatments
  - Blood cell counts (BCC) {to monitor adverse effects on normal tissues}
  - Access to **all** data recorded in the TOP trial data bases during the patient's treatment

- **Imaging Data (baseline: just before chemotherapy start)**
  - Ultrasound (DICOM)
  - Prospectively Somo-vu 3D US images
  - Digital mammography (DICOM) for some cases
  - PET and CT or MRI for certain cases (DICOM)
  - ***Three ellipsoidal axes of the tumour (obligatory)***.
  - Delineation of the necrotic, cystic, hemorrhagic and solid tumour regions on the tomograhic slices.

- **Histopathological and Molecular Data**
  - Histopathological profile (metastatic disease ?, tumour cell types etc.)
  - Photographs of HE histopathology slides (MIRAX scan system)
  - Topo IIα gene and protein, HER-2 gene, p53 gene, DNA array based gene expression profiling of the bioptic material
  - Estimated tumour cell responsiveness to the drugs under consideration

5.6.4.2 Process and Data description of the recommended administration of drug dose

- The above mentioned data is entered into the "Oncosimulator" which performs the simulation of the tumour response to chemotherapy.
- A rough estimation of the response of representative normal tissues is also made.
- As a result, the most probable outcome is predicted.

- Based on the oncosimulator prediction (mainly the expected tumour shrinkage) , the clinician judges whether or not the chemotherapy outcome would be beneficial to the patient under consideration by also taking into account his or her logic, expertise and even intuition.

- In case that the expected outcome is not judged as beneficial, the patient may undergo other therapeutic interventions. Otherwise, the chemotherapeutic scheme is applied to the real patient.

- The actual chemotherapy administration schedule is registered.

The following examinations are carried out during and after treatment:

➲ **During chemotherapy (prospectively)**

- Ultrasound imaging *after each CT cycle (and preferably on the 1st day of each week of the chemotherapeutic cycle)*
- Recording of the tumour 3 ellipsoidal axes

➲ **After completion of chemotherapy**

- Ultrasound (DICOM)
- Prospectively Somo-vu 3D US images
- Digital mammography (DICOM) for some cases
- PET and CT or MRI for certain cases (DICOM)
- **Three ellipsoidal axes of the tumour** (obligatory).
- Delineation of the necrotic, cystic, hemorrhagic and solid tumour regions on the tomograhic slices.
- Blood Cell Counts (BCC)

The predicted and the actual outcome are compared and if they are in significant contradiction an optimization and adaptation loop for the “Oncosimulator” is carried out, otherwise the current checking of the “Oncosimulator” is judged as favourable.

5.7 Scenario SC6: Molecular apocrine breast cancer

5.7.1 Background

Reference: Farmer P. et al., Identification of molecular apocrine breast tumours by microarray analysis, Oncogene, 2005 Jul 7;24(29):4660-71.

“Previous microarray studies on breast cancer identified multiple tumour classes, of which the most prominent, named luminal and basal, differ in expression of the oestrogen receptor alpha gene (ER). We report here the identification of a group of breast tumours with increased androgen signalling and a 'molecular apocrine' gene expression profile. Tumour samples from 49 patients with large operable or locally advanced breast cancers were tested on Affymetrix U133A gene expression microarrays. Principal components analysis and hierarchical clustering split the tumours into three groups: basal, luminal and a group we call molecular apocrine. All of the molecular apocrine tumours have strong apocrine features on histological examination (P=0.0002). The molecular apocrine group is androgen receptor (AR) positive and contains all of the ER-negative tumours outside the basal group.

Kolmogorov-Smirnov testing indicates that oestrogen signalling is most active in the luminal group, and androgen signalling is most active in the molecular apocrine group. ERBB2 amplification is commoner in the molecular apocrine than the other groups. Genes that best split the three groups were identified by Wilcoxon test. Correlation of the average expression profile of these genes in our data with the expression profile of individual tumours in four published breast cancer studies suggest that molecular apocrine tumours represent 8-14% of tumours in these studies. Our data show that it is possible with microarray data to divide mammary tumour cells into three groups based on steroid receptor activity: luminal (ER+ AR+), basal (ER- AR-) and molecular apocrine (ER- AR+)."

5.7.2 Scenario summary

This is a technology driven scenario. The objective is to use published results, use the available data and re-produce the results with the ACGT platform.

It is a "Microarray-data only" study, with one-color system (Affymetrix). No clinical data are provided.

Extra data required: two lists of probesets used to construct AR and ESR1 metagenes are provided on the ACGT/BSCW server.

5.7.3 Description of available data

Microarray data: from Phase III clinical trial EORTC 10994/BIG 00-01.

Chips: Affymetrix HG-U133A CEL files

NCBI GEO database accession number GSE1561

Data available from Oncogene website, relevant data copied under https://bscw.ercim.org/bscw/bscw.cgi/147172

5.7.4 Workflow

The forseen sequence of activities for the execution of the scenario are as follows:

- Introduce CEL files as clinical data in the ACGT database (e.g. BASE).
- Access the database from the ACGT analysis environment to load the data.
- Normalize data in CEL files with the method RMA in the R/BioConductor package affy.
- Reproduce the clustering of Figure 1b[7] (heat map, clustering simultaneously on genes and samples). Data were filtered as follows: only probesets with SD>0.5 (after log2 transformation) were considered.
- Assign labels (blue, green and red colors) to samples based on the above sample clustering.
- Perform a principal component analysis, and reproduce Figure 1a.
- Isolate genes from selected subclusters, together with their annotation (Fig. 1c).

[7] Figure 1b and others refer to figures as found in the relevant article

- Plot AR vs ESR1 gene expression (Fig 3a), and AR-metagene vs. ESR1-metagene expression (Fig 3b, using list of probesets composing metagenes AR.txt and ESR1.txt).
- Reproduce Kolmogorov-Smirnov (KS) tests8 in Figure 3.
- Gene Ontology analysis conducted with SIB program Io, to be reproduced in the context of ACGT analysis environment: http://www.io.isb-sib.ch/
- Optional: Comparisons with other studies appearing in paper.

5.8 Scenario SC7: van 't Veer study

5.8.1 Background

This again is a technology driven scenario, as published in the following reference.

Reference: van de Vijver et al., A Gene-Expression Signature as Predictor of Survival in Breast Cancer, New England Journal of Medicine, 347, 2002, 1999-2009

The following text provides information on the scope, methods, tools and results as reported in the reference:

"Background A more accurate means of prognostication in breast cancer will improve the selection of patients for adjuvant systemic therapy.

Methods Using microarray analysis to evaluate our previously established 70-gene prognosis profile, we classified a series of 295 consecutive patients with primary breast carcinomas as having a gene-expression signature associated with either a poor prognosis or a good prognosis. All patients had stage I or II breast cancer and were younger than 53 years old; 151 had lymph-node–negative disease, and 144 had lymph-node–positive disease. We evaluated the predictive power of the prognosis profile using univariable and multivariable statistical analyses.

Results Among the 295 patients, 180 had a poor-prognosis signature and 115 had a good-prognosis signature, and the mean (±SE) overall 10-year survival rates were 54.6±4.4 percent and 94.5±2.6 percent, respectively. At 10 years, the probability of remaining free of distant metastases was 50.6±4.5 percent in the group with a poor-prognosis signature and 85.2±4.3 percent in the group with a good-prognosis signature. The estimated hazard ratio for distant metastases in the group with a poor-prognosis signature, as compared with the group with the good-prognosis signature, was 5.1 (95 percent confidence interval, 2.9 to 9.0; P<0.001). This ratio remained significant when the groups were analyzed according to lymph-node status. Multivariable Cox regression analysis showed that the prognosis profile was a strong independent factor in predicting disease outcome.

Conclusions The gene-expression profile we studied is a more powerful predictor of the outcome of disease in young patients with breast cancer than standard systems based on clinical and histologic criteria."

[8] The procedure of Kolmogorov-Smirnov testing is also called « gene set enrichment analysis » (GSEA).

5.8.2 Goals

The scenario is a "Microarray with associated clinical data" study, with a two-color system (Agilent), data already background-corrected and normalized.

The main objective is: "the validation of a 70-gene predictor identified in a previous study by the same authors (van 't Veer et al., Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer, Nature, 415, 2002, 530)".

5.8.3 Description of available data

The original source of data can be found at Rosetta Inpharmatics web site: http://www.rii.com/publications/2002/nejm.html

Relevant data are copied under: https://bscw.ercim.org/bscw/bscw.cgi/147547

Access to Rosetta pre-processed list of 70 gene expression for the 295 samples (zip file for "table2") requires accepting and End-User License Agreement. This table is not included in the ERCIM repository as it can be reconstructed from the other files.

5.8.4 Workflow

The forseen sequence of activities for the execution of the scenario are as follows:

- Load sample data (expression ratios) in the environment.[9] [Extract expression data from the different files and create a single gene-expression/sample data matrix.]
- Associate clinical data with the samples.
- Retrieve data from 70-gene list in the 295 sample list.
- Compute correlation with 70-gene expression for each sample, using the values in the 70-genes file, and assign labels the sample according to their prognostic. (Use lymph-node status specific thresholds.)
- Reproduce survival curves (Kaplan-Meier) from Figure 2 of the article.
- Using binary coding for St-Gallen and NIH Consensus in the clinical data file, assign labels to samples and produce the corresponding survival curves.

5.8.5 Possible extensions to the set of clinical scenarios

A recommendation for extending the set of scenarios based on published data is provided by N. Graf, with Reference: B. Zirn et al., "Expression profiling of Wilms tumors reveal new candidate genes for different clinical parameters", Int. J. Cancer, 118, 2006, 1954-1962. This scenario would allow covering the spotted-array-specific features of the ACGT platform.

- Other scenarios based on published data can be included in the document namely:
- Upload and analysis of data from a two-color platform (Agilent, spotted-array).

[9] In an actual scenario, this step should be replaced by an operation of upload of raw data files for two-color system, followed by a normalization step.

- Multiplatform studies (e.g. combined microarray/PCR data), published data from UNC available.
- Extension of a study by using data published later (e.g. van de Vijver).

5.9 Scenario SC8: Antigen Characterisation Scenario

5.9.1 Background

The goal of this sectiom is to describe in sufficient detail the antigen characterization scenario co-developed between the University of Saarland “UoS” (an ACGT end user partner) and Biovista (a technology providing partner).

The scenario is based on actual current practices at the UoS as described in [GRATR] and discussed in the ACGT Technical Board meeting of Athens (11-13 July 2006).

5.9.2 Goals

A Clinical Trials investigator wishes to understand in depth the nature of the 50 antigens selected as a result of the screening process of the circa 30K proteins against the Wilms Tumor sera.

- The pattern of the identified antigens will contribute to answer key questions about the humoral immune response in Wilms tumor patients:
- Are Wilms tumors associated with frequent antibody response?
- Is there a complex and/or specific antibody response?
- Is this response associated with specific genetic features like gene amplifications or DNA losses?
- Do these immunogenic antigens share common features like specific sequence motives?
- Does the seroreactivity pattern allows early identification of Wilms tumors and also their histological subtypes?
- Does the seroreactivity pattern represents a prognostic marker for Wilms tumors in respect to chemotherapeutic response and / or outcome?

5.9.3 Benefits of solving this problem

Finding a typical pattern for nephroblastoma will help to make the correct diagnosis. This pattern will also be used as a tumor marker during follow-up. If different signatures between different histological subtypes are found, patients can be treated more individualized from the beginning according to their risk group.

5.9.4 Required and Available Data

The following data sources will be used for the present scenario:

- Medline database of 16.5 million abstracts
- List of 50 antigens supplied in text format (as a list or an Excel file) by the CT investigator.

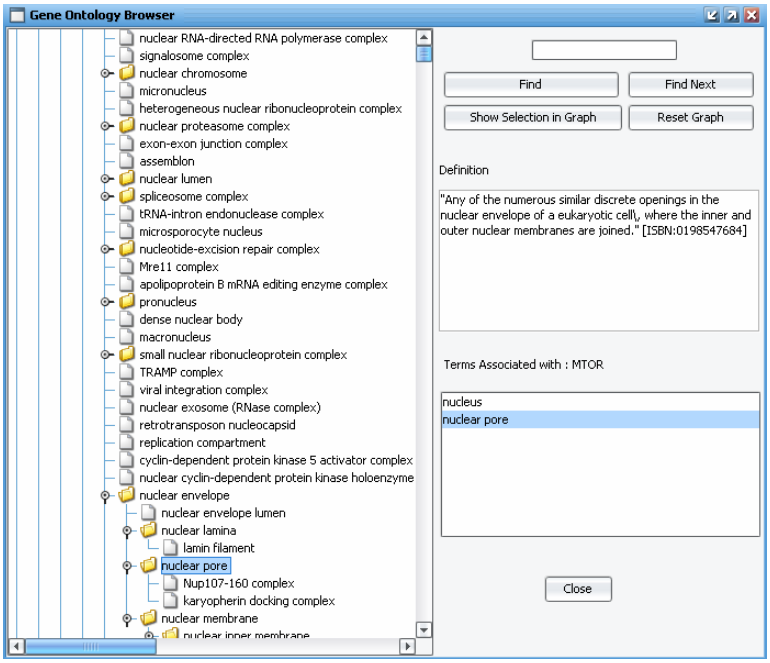
- List of URLs of on line gene resources
- Gene Ontology (GO) (http://www.geneontology.org/)
- Lists of proteins, diseases, pathways and a set of other molecular biology related concepts

5.9.5 Workflow

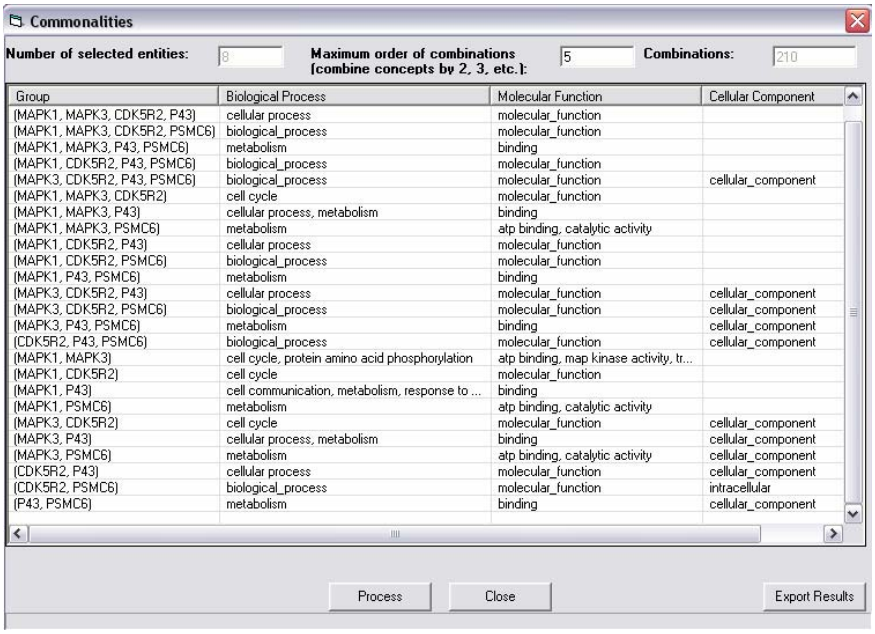
Two high-level goals are identified:

Goal 1*: to characterise and collect available information/knowledge about the 50 antigens.* In doing this a user needs to perform the following activities:

- Start with 50 proteins (antigens) given as a list of names (currently in list or Excel sheet format).
- Collect information from 3 distinct resources: online sources, ontologies and literature
- For each of the 50 antigens get information from one or more online resources such as KEGG and GO. This involves connecting to the appropriate web page, passing the name of the protein as a parameter, retrieving the returned information and presenting it in a format that is acceptable/usable by the user.
- For each of the 50 antigens see what ontological knowledge is available (using GO)
    - Find straight ontological information and display it in a format that looks something like the picture below:



    - Find commonalities (using BEA's "most recent common ancestor" algorithm) and display results in something like the format shown below
    - Find literature combining the 50 antigens or combinations thereof.
    - Take Wilms Tumor "gene signature" from the literature and compare this with the 50 antigens



G2: to generate and evaluate hypotheses based on information collected. In doing this a user needs to perform the following activities:

- Use the 50 antigens to find diseases for which they can be a predictor.
- Explore the identified diseases and understand the rankings by reviewing the supporting literature in each case (involves the reading of retrieved abstracts from the Medline database).

5.9.6 Technical Requirements for Scenario

The scenario described requires the following capabilities:

- Ability to input gene names (or other entities of interest such as adverse events, diseases etc) from existing sources including text, Excel files and others.
- Ability to correlate entities (e.g. gene names) to each other via their co-occurrence in the scientific literature (this can be generalized to any acceptable corpus such as the EPO life science patents).
- Ability to connect transparently to a variety of resources (e.g. KEGG) and perform basic bibliographic analyses.
- Ability to disambiguate search terms and filter results using a variety of criteria.
- Ability to perform the analytic tasks in acceptable times (order of minutes) and drill down to underlying literature in a transparent manner.

5.9.7 References

[GRATR] Norbert Graf,Technical Report SAR-ACGT-T2.1-final: Identification of Nephroblastoma antigens and determination of the seroreactivity, internal ACGT document

[BIOTR] ACGT Technical Report BVA-ACGT-6-1v : Biovista Technology Introduction, internal ACGT document

6 User Needs and Requirements

6.1 Introduction

Biomedical research has entered a new phase. The completion of the Human Genome Project sparked the development of many new tools for today's biomedical researcher to use in finding the mechanism behind disease.

Coupled with the sequencing and annotation of many model organisms, our ability to risk-stratify patients using a collection of phenotypic and genotypic information may come to fruition in the foreseeable future. Our ability to deduce correct dosage, drug efficacy, and provide pre-clinical intervention for at risk patients may become a reality.

While the goal is clear, the path to such discoveries has been fraught with roadblocks in terms of technical, scientific, and sociological challenges. The deluge of data that large-scale sequencing, transcriptomic and proteomic studies have produced to date is a case in point. In addition to the shear volume, data collected using a variety of laboratory technologies and techniques are often published without the background information (method of capture, sample preparation, statistical techniques applied) that is needed to reproduce results. In fact, a typical researcher spends as much time trying to understand the origins of a dataset as actually performing new analyses.

The situation is even more problematic in the clinical research domain, where ***data collection is still often performed on paper forms that differ from study to study,*** even when the same types of data are being collected. Rarely is a clinical biostatistician able to make good use of data collected on studies they were not directly involved with, largely due to ***incomplete or non-existent annotation and standardization of the information***.

This data problem has pushed the biological community to ***partition and compartmentalize*** their data for easy digestion and maintenance. While this approach worked in the past for simple systems containing a relatively small number of interactions, modeled by a small number of datasets, bioinformaticians are finding it difficult to model more complex systems. The simplicity and digestibility of the compartments described above have made it almost completely impossible to ***cross compartmental boundaries*** without consulting an expert.

To alleviate this burden, bioinformaticians are starting to apply sophisticated computational approaches in the areas of statistics, data mining, signal processing and artificial intelligence to discover relationships between such compartments. However, the community is quickly discovering glaring ***inconsistencies in language, methodology and computational models*** used to describe a particular organism, pathway, interaction, annotation, and so forth. The repercussions of the compartmental approach have produced a bottleneck in the road to discovery. To make this more concrete, computational approaches to data analysis and discovery typically rely on formalism in terms of syntax, context, and format in order to perform reproducible and consistent experiments (the backbone of hypothesis driven science). These formal definitions are severely lacking in the biological sciences. They will remain a burden to the process of biological discovery unless the biological community takes action.

The ACGT project aims to produce a software system capable of handling biomedical data that crosses compartments. The approach is to leverage existing technologies for the

standardization of data representation and semantics and couple them to a common data integration architecture. We propose a system capable of providing a "yellow pages" facility that identifies the provider, quality, provenance, language and methodology used to capture clinical, genomic, proteomic, and transcript data "on the fly" as well as services provided by the ACGT provider community to be used for the semantic integration of such multi-level biomedical data, for analyzing such data with the purpose of extracting new knowledge and all other tools and services required for setting up translational, multicentic clinical trials (Vitrual Organisation Set-up and Management) on the Grid.

Central to the approach of ACGT with respect to integration of heterogeneous, multilevel, biomedical databases is the ACGT Master Ontology (see ACGT DoW, pp 46-48). The ACGT Master Ontology uses industry standard techniques to define common biological constructs (objects) related to Clinical Trials and Cancer. By mapping local data sources to these common data objects, the process of (semantic) data integration can commence.

6.2 The ACGT users and stakeholders

The ACGT Project will ultimately impact anyone who is involved in post-genomic clinical research and clinical trials on Cancer. A closer analysis reveals that it appears to encompass a significant number of stakeholders. Stakeholders are defined as:

> *"An enterprise, organization, or individual having an interest or a stake in the outcome of the engineering of a system."*

This should not be a surprise given the importance of the domain of application, i.e. post-genomic cancer research form a large number of organisations. Indeed, the breadth of the project spans the universe of post-genomic, individualised medicine and information technologies, while the classes of stakeholders are quite large and diverse with respect to areas of interest, depth and domains of expertise, and desire for involvement.

Key repsresentative users are members of the ACGT Consortium, whereas others will be drawn into the ACGT user community as the project progreses.

Different user "levels" need to be considered. From the viewpoint of access to the components of the ACGT infrastructure, we identify the following actors:

- ***System administrator*** - With permission to manage the internals of the system.
- ***Service developers*** - they provide new applications. A controlled and supervised registry of such new services should be provided, including a version control mechanism.
- ***Registered (authorized) user*** - We can see two roles regarding data:
  - ***Data producer***. This could be a clinician who "owns" the data collected in the context of his/her clinical study. The data owner should be the only one authorized to modify his data. It is needed to identify the ownership of data submitted to the ACGT system.
  - ***Data consumer***. This could be any clinician who does not participate in a trial but is interested on getting some data from it. This role involves implementing

a policy of access to trial data. Data storage, data distribution and microarray data have to be in accordance to ethical and legal issues.

Regarding end-users of the ACGT infrastructure one can further distinguish:

- *Clinicians*
- *Biomedical researchers*
- *Data miners*
- *Patients*

Clinicians typically want to apply very standardized procedures. For instance, given a microarray result for a patient, they want to apply well known predictors combining clinical and microarray data (e.g. St-Gallen or NIH consensus in breast cancer) and get the results in the clearest possible fashion. Simple interfaces (e.g. web-based) should usually be sufficient for this category of users.

The category of ***biomedical researcher****s* includes typically PhD students, post-docs and senior scientists working in a clinical or research environment. Such users typically want to have a greater flexibility in the visualization of their data, and to have the opportunity to use various statistical tests and tools, but usually on a single set of clinical data. Interactivity in the tools is essential for these users (e.g. ability to click on a gene to get information about it, to filter their data based on selecting one branch of a cluster, to visualize gene expression as a function of the position on the genome [which typically requires zooming capabilities], to build Venn diagrams of gene lists found significant under different conditions, etc…). To some extend, these users will develop automatic methods (workflows) to complete their jobs. These WFs will be used by inexperienced users

The category of ***data miners*** overlaps somewhat with the previous category (e.g. some biomedical researchers may be willing to combine their findings with those of other studies). Data miners are typically biostatisticians willing to extract knowledge from a study using statistical tools they just developed or from a combination of studies or both. Meta-analysis tools would be of interest for that category of users. The tools used are mainly programs and scripts (e.g. in R), with less strong need for interactive graphics.

The role of the ***patient*** would be limited to access his/her clinical record, and to obtain various summaries of the findings associated to his/her clinical data. From the viewpoint of the ACGT infrastructure patients may be considered as clinicians with an access limited to a single subset of data (their own). As the ACGT infrastructure should essentially be a platform for improving clinical data knowledge mining, we recommend that the focus of the development effort to be put on the first three categories of users. Patients access to ACGT should be kept for a later phase of development of the infrastructure.

The key stakeholder groupswe see in ACGT are briefly described in the following sub-sections.

6.2.1 Cancer Research Organisations

The Cancer research community depends on data located in multiple disparate data sources around the world. In most cases, these data sources do not utilize the same underlying data structure. However, given these tremendous barriers to their research and discovery efforts, the cancer research community has been able to perform a wide range of studies to elucidate the mechanisms and treatment of cancer.

Most participants in the cancer community have access to the internet and use protocols such as file transfer protocol (FTP) and hypertext transfer protocol (HTTP) to access disparate data sources. As the community grows, many participants have become blind to the seemingly endless number of new datasets, methods and tools published on a daily basis.

The project intends to engage a large number of such organisations in its future "Requirement Engineering" activities. Some of the European Societies or organisations that play an important role in Cancer Research are listed in Appendix 3 in alphabetical order.

The following steps will be undertaken to promote ACGT:

1. All of the above mentioned societies and organisations will be officially contacted by ACGT asking for an official meeting with representatives at their annual or biannual meetings, conferences or congresses. The purpose of such a meeting is:
    - To introduce ACGT
    - To explain what can be expected from ACGT
    - To define possible areas of collaboration and
    - To ask for possibilities for affiliation
2. There will be active participation by giving a general ACGT presentation at the National Cancer Meetings, Conferences or Congresses by National representatives of ACGT.
3. There will be active participation at International annual or biannual congresses by members of ACGT to present ACGT to the participants of these meetings. Most important are the following International conferences:
    - European Cancer Conference (ECCO)
        - The next conference will be held in:
            - Barcelona, Spain, September 23-27, 2007,

              http://www.fecs.be/emc.asp?pageId=1228&Type=P
    - Congress of the International Society of Paediatric Oncology (SIOP)
        - The next congresses will be held in:
            - Geneva, Switzerland, September 17-21, 2006,

              http://www.siop.nl/siop2006
            - Mumbai, India, October 30- November 3, 2007,

              http://www.siop2007.in/
    - Annual Meeting of the American Association for Cancer Research (AACR)
        - The next meetings will be held in:
            - Los Angeles, CA, April 14-18, 2007,

              http://www.aacr.org/page6899.aspx
            - San Diego, CA, April 12-16, 2008

              http://www.aacr.org/page6901.aspx
4. A Management Board Meeting (MB) will be organized at the ECCO conference in Barcelona in 2007. The organizers from the Federation of European Cancer Societies (FECS) will be asked to officially put the ACGT MB Meeting into the program of the conference. This seems worthwhile to do, because one of the topics of the ECCO meeting in 2007 is the translation of science into better clinical practice.

5. The different National Cancer Research Centers and National Cancer Institutes of enrolled partners in Europe and Japan will be contacted by members of ACGT. A list with contact details is given in Appendix 4.

6. The Homepage of the ACGT Project will be used to spread information of the project to Cancer Research Organisations.

6.2.2 Researchers and Scientists involved in post-genomic research

The lack of a unifying architecture has proven to be a major roadblock to a researcher's ability to mine different databases. Most critically, however, even within a single laboratory, researchers have difficulty integrating data from different technologies because of a lack of common standards and other technological and medico-legal and ethical issues. As a result, very few cross-site studies and clinical trials are performed and in most cases it isn't possible to seamlessly integrate multi-level data (from the molecular to the organ, individual and population levels).

Moreover, clinicians or molecular biologists often find it hard to exploit each other's expertise due to the absence of a cooperative environment which enables the sharing of data, resources or tools for comparing results and experiments, and a uniform platform supporting the seamless integration and analysis of disease-related data at all levels and simulating the mechanisms of disease evolution.

The project also intends to engage a large number of such researchers and scientists in its future "Requirement Engineering" activities. Two main processes will be undertaken to achieve this goal.

- The questionnaire, that was developed in task 2.2, provides names of researchers and scientists, to whom a similar questionnaire will be send. The number of names will be increased by adding representatives from the different Cancer Research centers and Societies as well as from IT domains to this list. Such a questionnaire will open the communication with researchers and scientists in different fields and areas, that are important for the project.

- The Homepage of the ACGT project provides the possibility of getting involved as an associated member. With such associated members future requirements will be discussed. This will be possible via the homepage by integrating a Wiki for this purpose.

6.2.3 Technology Suppliers

Technology stakeholders are interested in designing, building and integrating, that would effectively become a part of the implementation of the integrated ACGT technology platform.

These individuals would be adopters of the architecture specifications and associated standards. As such, vendors provide a different view of the platform's requirements – one clearly focused on the implementation and long term use of the architecture. The purpose of engagement with vendors is to gain acceptance of the project concepts from vendors and suppliers directly designing, building, and supplying equipment. Many vendors already participate in standards bodies and this project must be presented in the context of building upon existing standards development work.

Such manufacturers are likely to have substantial technical input that must be sought as the technical requirements are being established.

6.2.4 Patients and Patient Organisations

In greater Europe, there are 2.8 million new cases of cancer every year, and 1.7 million die from the disease every year. In the EU 25 alone, 4.5 European citizens have had or have cancer. There are 2 million new cases every year, and 1.16 million die from it every year (2004).

The growing cancer burden in Europe can be illustrated in many other ways, each telling of individual suffering and loss, but also of a need to act collectively to eliminate this scourge:

- 1 in 3 men and 1 in 4 women will be directly affected by cancer in the course of the first 75 years of their life.
- 3000 people die from cancer every day in Europe.
- Cancer is the **second main cause of death** in Europe, after circulatory diseases.

By complementing public actions towards better health and by collaborating with local and European authorities to fight cancer it is of utmost importance to coordinate this lobbying activities in close cooperation with patients and patient cancer organisations. Today for nearly every cancer a patient organisation is known. In the European context patients' coalitions did form Cancer leagues, to provide support to cancer patients and their relatives, and to improve the quality of treatments.

The **Association of European Cancer Leagues** (ECL) (http://www.europeancancerleagues.org/) is a federation of national and regional Cancer Leagues, made of either patients' coalitions or of cancer control professionals. The objectives of the association are to improve communication, to promote, enhance and co-ordinate collaboration between European leagues/societies and to foster fruitful activities between European cancer leagues and organisations, in order to reduce the growing cancer burden in Europe.

ECL is located in Brussels and is a non-for-profit association (asbl; association sans but lucratif), under Belgian law. ECL was created in 1980, and consists of 31 members today, located all over extended Europe (See Appendix 5 for a listing and details of members).

Regarding Paediatric Cancer the **International Confederation of Childhood Cancer Parent Organisations** (ICCCPO) was founded in May 1994 in Valencia, Spain. ICCCPO is a worldwide network of organisations of parents of children with cancer. The mission of ICCPO is to share information and experience in order to improve access to the best possible care for children with cancer everywhere in the world.

ECL as well as ICCCPO will be contacted by ACGT asking for collaboration and for providing names of persons, who will attend Management Board Meetings of ACGT. These representatives of the patient organisations should especially be enrolled in the discussion of legal and ethical issues. The goal is to strengthen the collaboration between patients and ACGT to foster fruitful activities in order to bring basic research faster into clinical practice.

6.2.5 Regulatory Agencies

Regulators and auditors have an interest in ensuring that the ACGT systems meet their reliability, performance, market, and financial obligations.

The purpose of this engagement is to assist regulatory commissions in understanding the nature and need for a project that develops an industry-wide architecture and the problems that arise from lack thereof.

6.2.6 Standards Bodies

The purpose of engagement with these groups is to gain acceptance and future standardization of the concepts. Engagement with standards groups such as the IEEE, GGF, and others is planned.

Also, clinical trials related standards bodies, such as the International Conference on Harmonisation (ICH) which is a clinical trials related standard. The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) is a unique project that brings together the regulatory authorities of Europe, Japan and the United States and experts from the pharmaceutical industry in the three regions to discuss scientific and technical aspects of product registration.

The purpose is to make recommendations on ways to achieve greater harmonisation in the interpretation and application of technical guidelines and requirements for product registration in order to reduce or obviate the need to duplicate the testing carried out during the research and development of new medicines.

The objective of such harmonisation is a more economical use of human, animal and material resources, and the elimination of unnecessary delay in the global development and availability of new medicines whilst maintaining safeguards on quality, safety and efficacy, and regulatory obligations to protect public health.

Some issues that standards bodies and industry consortia can expand upon for the ACGT team include:

- Drivers to harmonize standards in progress
- Needs to address standards integration on enterprise and industry levels
- Integration of standards initiatives and past work
- Drivers and needs to establish interworkability testing and other formal methods of integration
- Integration of architecture methods with standards development activities

6.2.7 Market Participants

Market participants, as part of the User Community, have unique communication requirements, reflecting their market-driven needs for timeliness, availability, and security of many different types of information. These requirements are not always met, particularly as the market environment, policies, and capabilities change over time.

6.3 Requirements Analysis

6.3.1 Global requirements

In order to provide some assistance to meeting the objectives, it has been deemed useful to enumerate a set of requirements and assumptions contextualizing the challenges of the *bioinformatics* side of the ACGT project. Those requirements are the following:

6.3.1.1 Data repositories

1. **Massive data production**. Nowadays, many technological breakthroughs such as high-throughput screening, sequencing and gene-expression monitoring technologies have nurture the 'omics' revolution enabling a massive production of data; from raw DNA sequence composition, tri-dimensional structural information and gene-expression data about the dynamic behavior of genes under certain experimental conditions. Molecular data production has shifted from one-gene experiments to whole genome scale.

2. **Diversity, heterogeneity and dispersion**. Although it is a commonplace statement that the volume of data in molecular biology is growing at exponential rates, nonetheless, the key characteristic of biological data is not so much its volume, but its diversity, heterogeneity and dispersion, which make this plethora of interrelated information difficult to use.

    a. Diversity is related to the many different but related biological data types and the number of different repositories regarding to each data type (sequences, structures, gene-expression data, etc)

    b. Heterogeneity is related to the different format in which the data are available (most of them as plain text, images, video, etc). The various repositories commonly have different data formats, access mechanisms and user interfaces, which is in addition complicated by the different terminologies used.

    c. Dispersion is related to the geographical location of data servers

3. **Dynamic data production**. Updates to local databases occur frequently, so the newest data should be accessible, too (coarsely, four releases per year). In other words, the contents of the local databases may be autonomously and locally maintained.

4. **Data mirroring**. ACGT system should ensure data availability by replication (data mirroring). This situation needs database content synchronization in the different mirrors at any time, which has to be ensured.

5. **Change of repositories**. In some cases the schemas of local repositories change (in some cases even two or three times per year!). The ACGT database system should be designed and maintained to meet local needs. Changes of the repositories must guarantee data integrity and should be made possible independently of the integrated database structure.

6. **Biomedical database**. The situation is even worse in the case of Medical Informatics in which, despite the efforts and advances in the area of standards, a single model for biomedical databases will not emerge, at least in the short term.

6.3.1.2 Bioinformatics and other analytical service requirements

From the service provision perspective, bioinformatics is a highly dynamic environment in which services arise dynamically when new algorithms are implemented and deployed in the form of programs running on a server connected to Internet

A broad range of processes are available (I/O and/or CPU bounded). Some processes demand only few seconds to be completed but involve large input/output needs (similarity database searching like Blast, Fasta, etc.); or on the other hand, they demand long CPU-times to solve the problem (i.e. molecular dynamics).

Under these considerations some basic principles should guide the design of the ACGT platform:

- Component-based (instead of monolithic approaches). Traditionally, bioinformatics solutions are based on the combination of different tools and data (from different sites) to produce a global insight into biological functions. In this regard, a small piece of code specifically devoted to solve a given task can be combined and re-used in different ways to produce the desired results.

- Bioinformatics research is mainly performed in a web-based working environment in which users want to use services without regard for details such as localization, implementation aspects, etc. Their only interest lies in the results they can provide or get.

- Facilitate the incorporation of new high quality and high demanded services (with controlled registry). A balance between the desirable expansion of the catalogue of services and the quality they exhibit must be maintained. Only this can ensure the success of the global offer. As new interesting services are multiplying dynamically some rules for their integrated use must be observed.

- Uniform deployment of services should be provided: due to diversity, dispersion and service heterogeneity, automatic procedures should be available to generate corporate –yet customizable- user interfaces, help system, etc. in a standard, efficient and friendly way.

- The system should be self contained, i.e. the necessary tools for entering data into the system (in the user's own terms, whatever the design), or visualizing results, should be provided by the platform.

- Service-discovering procedures should be supplied to promote automatic connection and service pipelining. To allow this, a common scheme of reference (a semantic description of data) should be established by means of a controlled registry of services.

- Platform scalability should be provided by means of intelligent resource management to allow increment on the number of services available and the tools provided.

- Robustness of the platform should be ensured by service and data replication at different sites.

- Data and service coherence should be maintained independently of what, where, when, how and by whom data or services are being used.

- Enabling standard storage, secure access and exploitation of proprietary data should help reduce overhead costs.

- Efficient use of computational resources should be achieved by intelligent computational load distribution among the different instances of the same service (i.e the "quality" of the service can be measured as function of the response time, and instant computational load should be taken into account.

- Connectivity of the platform to external machine-to-machine procedures should be considered. Apart from the traditional interactive use of the system, it is of interest to provide a capability of automatic request for services without going through the GUI. This would allow the inclusion of services in batch calls for large tasks or to proceed with workflows or for re-engineering services by coupling simple code-pieces.

- Interactive visualization. This is a key aspect on the capabilities of the platform to be designed and –as it is supposed to be oriented- a web-based environment, which is not especially suitable for this kind of task. So especial effort must be devoted to address this issue in an efficient way.

- Workflows as job-solving paradigm. Combining simple tasks services will allow the dynamic and flexible inter-connection of services thus creating complex distributed bioinformatics machines. This will expand the functionality and will enable the easy incorporation of new procedures to customize the system for specific concerns.

- Ethical and legal issues have to be considered at any time and in all services that store, distribute or integrate personal data or data that are connected to patients. This is also a main aspect in dealing with biomaterial data.

6.3.2 Technical requirements resulting from the ACGT scenarios

In a clinical data analysis one typically starts with raw, uncurated data to reach a high-level knowledge after analysis. This process typically relies on:

- Ability to load data from different technological platforms (Affymetrix, Agilent, custom spotted arrays, PCR data) and/or preprocessed data table.
- Ability to associate clinical data to samples. Including ability to handle missing values, and ability to track sample/patient/array associations.
- Ability to address the required, by legislative and ethical requirements, anonymization and privacy requirements.
- Ability to apply various low-level[10] analytical tools (e.g. normalization, background-subtraction, QC of raw data) to check the quality of the raw data and ensure that they can be compared.
- Ability to store information about well known diagnostic/prognostic tools (*e.g.* in breast cancer context: St-Gallen Consensus, NIH Consensus, NPI, 70-genes criteria).

[10] In the present context "low-level analysis" implies the dependence of the analysis tools on the choice of technology (e.g. one- vs. two-color microarray technologies), while "high-level analysis" implies independence from the technological platform; again in microarray context, once the expression matrix is obtained for all genes, tools such as PCA, clustering, etc… can be applied independently from the platform.

- Ability to use high-level analytical tools for exploratory and validation purpose (e.g. PCA, clustering, logistic regression, cross-validation procedures, identification of genomic duplications and/or deletions, etc…)
- Ability to integrate known information/annotation in the analysis, such as pathways, literature information, gene ontology, etc. Such information must be in a format usable in the analysis, e.g. as filtering criteria.
- Ability to use high-quality visualization tools (e.g. mouse-moveable 3D scatterplots, pathways, visualization of expression in genomic context …), and to easily produce figures "ready for publication".
- Ability to search for relevant publications.
- Ability to "script" analyses such that they can be easily repeated and/or extended, i.e. defininion of elementatry workflows and their re-use.
- Some scenarios are "basic" scenarios in the sence that the required functionality is also essential in other, higher level and more complex, scenarios. This fact reveals the requirement to oechestrate "elementary workflows" and their invocation and integration into more complex scientific workflows.

Each of these points requires specific tools which either already exist or which need to be developed nearly independently from the others (once interfaces are defined).

6.3.3 Requirements and Use Cases for the ACGT Grid

A number of administrative use cases will be required and some have already been identified[11]. These use cases do not directly follow from the user scenarios but rather result from re-using experiences of other related Grid projects (e.g. caBIG, myGRID, IntelliGrid,etc). Some of these use cases are:

- **Virtual Organisation setup and Management:** This use case describes the process for the creation of an ACGT Virtual Organisation and its management.

- **Install ACGT Grid wizard:** This use case describes the ACGT Grid installation on a user's computer.

- **Startup:** This use case describes the ACGT startup process triggered by either the operating system or the ACGT user. During the ACGT installation, the user can select to start the service automatically or manually. When automatic is selected, ACGT is launched when the system boots up. When manual is selected, ACGT is launched by the user from the command line.

- **Shutdown:** This use case describes the ACGT shutdown process is triggered by either the operating system or a user. The functionality will be command line.

- **Configure Data Source and Domain**: This use case describes how to configure the ACGT Grid either locally or at the site level. The use case includes configuring new data source and new domains to a local ACGT.

[11] Similar use cases have been identified in caBIG, myGRID and a number of other Grid projects.

- **Setup Privileges:** This use case describes the configuration and functionality of the security component involved in ACGT Grid. The security identifies who can access what data.

- **Login Admin Tool:** This use case describes the ACGT Grid login admin tool graphical user interface (GUI) process.

6.3.4 The ACGT Master Ontology

The development of the ACGT master ontology involves the analysis of (a) the ontological needs of the ACGT clinical scenarios and (b) the ontological foundations and coverage of the existing terminologies and ontologies in the biomedical domain. Based on the analysis of (a) and (b), it will become possible to craft an ontology that is able to function as a semantic mediator between all systems to be integrated. All the classes and relationships of the ACGT master ontology must be given a formal definition.

6.3.5 The ACGT Mediator

In enabling integrated acces to heterogeneous and distributed biomedical databases (local, or public) necessitates the development and use of a mediating component (in conjuction with the ACGT Master Ontology).

The *mediator* is a service enabling 'ontology-based" integration of heterogeneous data sources (which can be ACGT databases, external databases, web sources, web data services) by providing a virtual view of all this data.

Users (including ACGT tools or services) asking queries to the mediator system do not have to know about data source location, schemas, or access methods, because the system will present one shared mediator ontology (*ACGT **master ontology***) to the user and users will ask their queries in terms of it.

6.3.6 Requirements and Use Cases for data access and analytical services

A range of requirements and use cases have been identified following a top-down approach, i.e. by analysing the overall vision of the project and taking into account results of other similar projects and initiatives world wide technological stakeholders have identified several needs and requirements as listed below.

6.3.6.1 Mapping and Advertisement

Advertise Analytical Service: This use case describes how ACGT Grid at a given site publishes data services. That is, what domain objects is it going to serve and based on what criteria.

Advertise Data Source: This use case describes how data sources at a given site are published in the ACGT Grid. That is, what domain objects they expose, their metadata descriptors, etc.

Configure Object to Ontology Mappings: This use case describes how to map Objects fromthe Master Ontology with a local Data Source at the meta-data level. The mapping process is between one object attribute from the object model fromthe Master Ontology with one field-table from the data source. After the actor maps the data, the mappings persist and

are subsequently used during local data retrieval. Object to data mapping can be launched from the installation process or from an appropriate administration tool.

6.3.6.2 Query and Discovery

Query Data and Discovery Service: This use case allows the user to retrieve data from multiple Data Sources. These Data sources are available through Advertisements/Services.

The actor should have the option to query the system in a number of different ways, e.g. request data sources (local and remote), and request attributes, or request data. Using this API, the ACGT client will be able to discover all Data Sources and the metadata that they support.

Discovery is the process of one peer searching for another peer, in the same peer group, that contains the desired content

6.3.7 Requirement for Semantically Discoverable Services and Metadata

As ACGT aims to connect data and tools from many disparate cancer centers and other organisations developing and contributing analytical tools, a critical requirement of its infrastructure is that it supports the ability of researchers to discover these resources.

ACGT will enable this ability by taking advantage of rich semantic descriptions of data models and services that are available. Each service is required to describe itself using the ACGT service metadata. When a service is connected to the ACGT Grid, it must register its availability and service metadata with a central indexing registry service (***Index Service or Metadata Registry***). This service can be thought of as the "**yellow pages**" and "**white pages**" of ACGT. A researcher can then discover services of interest by looking them up in this registry.

ACGT will provide a set of high-level APIs and user applications for performing this lookup which greatly facilitates the process. As the Index Service contains the service metadata of all the currently advertised and available services in ACGT, the expressivity of service discovery scenarios is limited only by the expressivity of the service metadata.

This service metadata should contain information about the service-providing cancer center, such as the point of contact and the institution's name. Extending beyond this generic metadata are two standards that are specialized for the two types of community-provided services: Data Services and Analytical Services. The Data Service Metadata details the Objects (as represented in the ACGT Master Ontology) being exposed by the service. Similarly, the Analytical Service Metadata details the Objects involved and defines the operations or methods the service provides. The input parameters and output of the operations are defined by referencing the appropriate Object definition. In this way, both the data and analytical services fully define the domain objects they expose by referencing the Objects registered in Master Ontology.

6.3.8 Requirements for Workflows

The ACGT master ontology, along with additional service/workflow metadata and ontologies will also be used for annotating services and ready made workflows (involved in wet lab experiments). Service and workflow annotations provide information regarding the service interface, functionality, provider, quality of service, etc. Annotated services and workflows are registered in the service/workflow registry, organized in classes.

Based on these annotations, and assisted by the service and workflow discovery module, the user can semi-automatically compose new workflows. In a workflow, data and parameters are given as input by the user to the top-level services. Then, their output (possibly combined) is given as input to the next level services, and so on, until the final result is derived by the bottom-level service.

The workflow composition component should ensure that the output-input interfaces of the dependent services match. Once a workflow is composed, the user can execute it and store the result in the *Wet Lab DB*, annotated with (i) metadata and ontology terms describing the result, and (ii) provenance information (service invocation sequence, origin of data, dates, etc.), and based on the provenance metadata and ontology.

The use of ontologies, metadata, wrappers and workflows in wet lab experiments is shown in the figure below.

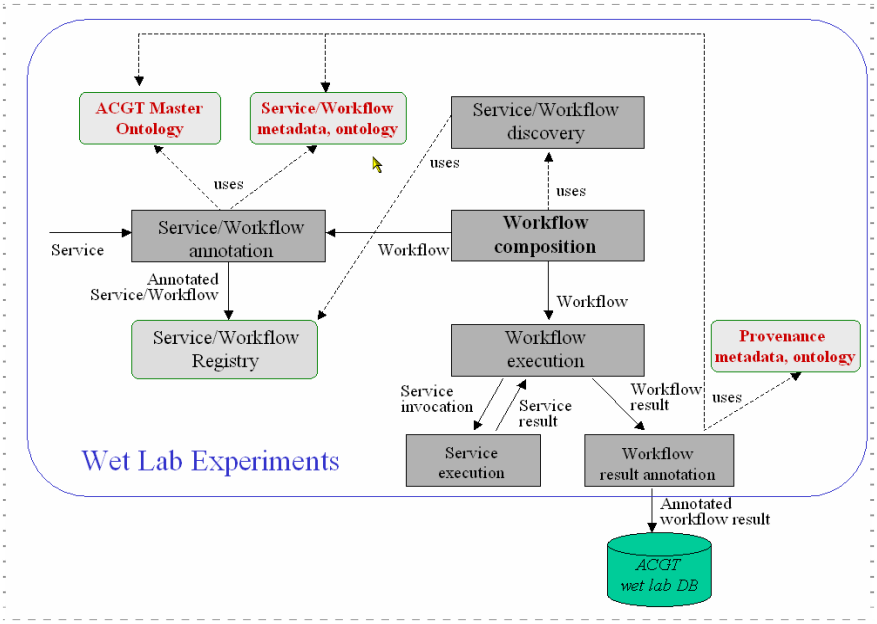


Figure 9: The use of ontologies and metadata in ACGT

6.3.8.1 Using the right level of abstraction

Workflow users will typically want to use remote services at different levels of abstraction depending on what they want to do. Some users will want to interact intimately with a specific service to tweak parameters that determine the detailed nature of the results or to tune performance. Other users will wish to be abstracted from these details since they are more concerned with the overall orchestration of several services into a high-level flow and hence want detailed workflows and specific invocation methods to be 'wrapped' up and delivered in an easy to use form.

6.3.8.2 Workflow lifecycle

Use of workflow as part of a scientific endeavour requires support for the workflow lifecycle. For example, a particular workflow will typically be authored, enacted, validated and modified in an iterative cycle. Whole workflows, or workflow fragments, will be published and shared so that others can use or learn from them, which in turn involves a process of annotation, discovery and personalisation. Therefore, workflow authoring, versioning, and scientific validation will be a key requirement from the ACGT technology platform.

6.3.8.3 Semantic description of workflows

The required workflows and resources to be accessed and orchestrated for a particular post-genomic study will not necessarily be known a-priori. Specification at a semantic level of the resources and activities required should allow discovery of suitable resources and workflows in a way that is abstracted from the syntactic details of data formats or invocation mechanisms. The use of explicit and machine-readable semantics for the inputs, outputs, and function of a workflow will increase the ability to share workflows since it allows workflows to be indexed, browsed, and searched according to their overall purpose rather than detailed syntax, data formats or service bindings.

6.3.8.4 Workflow provenance

Provenance is a fancy word that refers to an object's history, or who owned the object, when and where. It is to an object what a deed trail is to a piece of land. Provenance is the origin or source from which anything comes. The term is often used in the sense of place and time of manufacture, production or discovery. Comparative techniques, expert opinion, written and verbal records and the results of tests are often used to help establish provenance.

Use of workflows as part of scientific activity often require provenance data to be kept about the activities performed during workflow execution (recording of intermediate data sets, details of the specific service providers used, versions of data and tools involved, interventions were made by the user). Provenance support is needed in the workflow language (so that the required level of provenance can be specified); the systems used to enact the workflows (so that the specified provenance data is generated during execution); and data stores (so provenance data can be store d and subsequently retrieved).

6.3.9 Service Registry and Metadata

The system to be built for ACGT is expected to incorporate a multiple and diverse collection of databases and tools. Preparing specific interfaces to access such long list of services could represent an enormous effort. In our opinion, much of the strength of an integration system lies in the ability to combine different applications over a set of data. For this, the description of input/output objects must be coordinated and standardized by means of an object-ontology in such a way that services can inter-connect, understanding what the output from a previous process is. In the same way, services characteristics, i.e. metadata (aim, location, authority, help documentation, parameters, etc…) must be defined and stored in such a way that automatic engines can access this information to automatically build-up services interfaces, launch and manage the user request.

6.4 Analytical tools to be integrated in the ACGT environment.

The preliminary list of tools that have been identified as required and should be considered for inclusion in the ACGT environment is given in the “Tools description” section below. First priority should be given to the following tools:

- Heatmaps and Interactive hierarchical clustering tool focusing on two-way clustering (Possibly other clustering tools).
- Interactive representation of data in genomic context
- Pathway visualization

- "Gene" information visualization tool (problem: requires connection to "external" knowledge bases).

Other tools are as important in their functionality but could be developed later. An extended list of such tools is included in Appendix 2, which contains the result of an internal inquiry regarding available tools used by project participants and of other tools that might prove necessary to integrate them into the ACGT architecture, as additional requirements and user scenarios are developed.

6.4.1 Tools description

6.4.1.1 Data inspection tool

Before going into higher level of analysis, data are usually inspected visually, e.g. to identify outliers. Clinical data are usually represented in matrix form, values (e.g. microarray expression value) vs samples.

6.4.1.1.1 Detailed specifications

Basic actions:

- Histograms, distribution and boxplots for all or a subset of samples
- Toggle-button between log2/non-log2 transformation
- For microarrays: background subtraction, normalization, visualization of raw data according to the geometry of the chip, QC procedures (such as R's AffyPLM)

6.4.1.2 "Gene" information visualization window

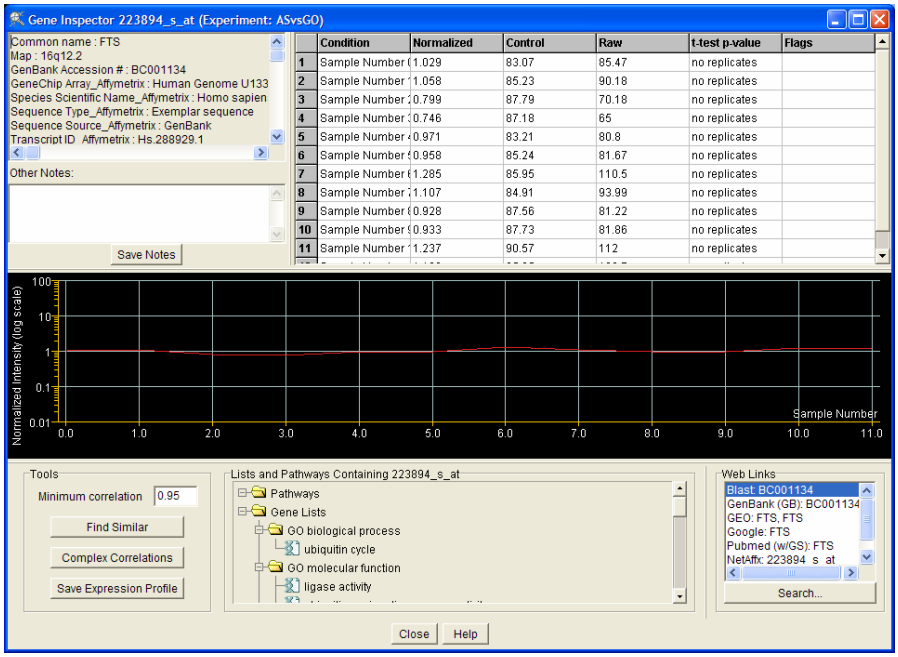
A generic tool to be developed is a feature data visualization tool. This tool can be reused every time the user clicks on a graph to obtain information about a gene.

6.4.1.2.1 Detailed specifications

Given a feature (e.g. spot in a microarray experiment, or a gene in proteomics), its annotation and the entire set of data in a clinical trial (or a subset of it), here are the basic actions to be realized in the "gene" information window:

- Provide the gene annotation
- View the gene expression in other samples in the trial
- Identify pathways and other lists containing the gene
- Find other genes with similar expression pattern
- Link to external databases (Ensembl genome viewer, NCBI, etc…)

Example: GeneSpring's Gene Inspector



6.4.1.3 Heatmaps

Heatmaps are simple plots to represent data of matrix type. Static heatmaps can be produced with R. Dynamic heatmaps would allow investigating the data in the matrix in a more efficient way.

6.4.1.3.1 Detailed specifications

A simple heatmap object consists of a matrix of numbers (possibly with missing values) with column and row names. A user-definable color scale is associated to the object (including a special color for missing data).

Basic actions:

- Clicking on a part of the graph would yield the column and row indices and names of the heatmap cell, as well as the value in the cell. (Cell selection.)
- Clicking-and-dragging over an area of the graph, yields vectors of column and row indices and names, as well as the associated matrix of values. (Area selection.)
- Zooming in and out

Extended actions:

- Double-clicking a row or column header opens a window containing a description of the associated gene/sample/protein…
- Clicking on a row or column header selects that row or column.
- Control-click allows multiple selections.
- Right-clicking once a selection is done opens a contextual menu, allowing several actions such as: Hiding the selection, Memorize the selection in a list to be returned once the plot is closed, etc…
- Axes and color scales should be user definable.

6.4.1.4 Interactive hierarchical clustering tool

Hierarchical clustering is one of the basic tools in unsupervised genomic analysis. Both one- and two-way clustering is used. A two-way clustering representation is essentially two one-way clusters with a heatmap showing the clustered data matrix. Existing tools include: Cluster/TreeView, GeneSpring, SpotFire.

6.4.1.4.1 Detailed specifications

One-way clustering.

Basic actions:

- Selecting metric and clustering algorithm.
- Selecting a set of branches of the cluster returns the identifiers of the selected genes/samples, as well as a structure describing the selected nodes in the tree graph.
- The cluster representation must allow extra parameters associated with the branches to be visualized.
- A vector of extra annotation can be provided and displayed.

Extended actions:

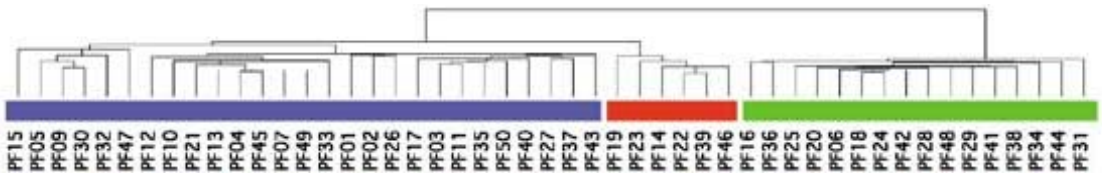
- Double-clicking on the leaf of a tree opens a gene/sample/protein… description window.

Two-way clustering.

Basic actions:

- On clusters: same as one-way clustering
- On heatmap: same as above

Example: one-way hierarchical cluster with extra information associated to leafs.



6.4.1.5 k-means clustering and self-organizing maps.

k-means and SOM are tools frequently used in clinical data analyses. Features similar to hierarchical clustering should be provided. Existing tools include: Gedi, GeneSpring, SpotFire.

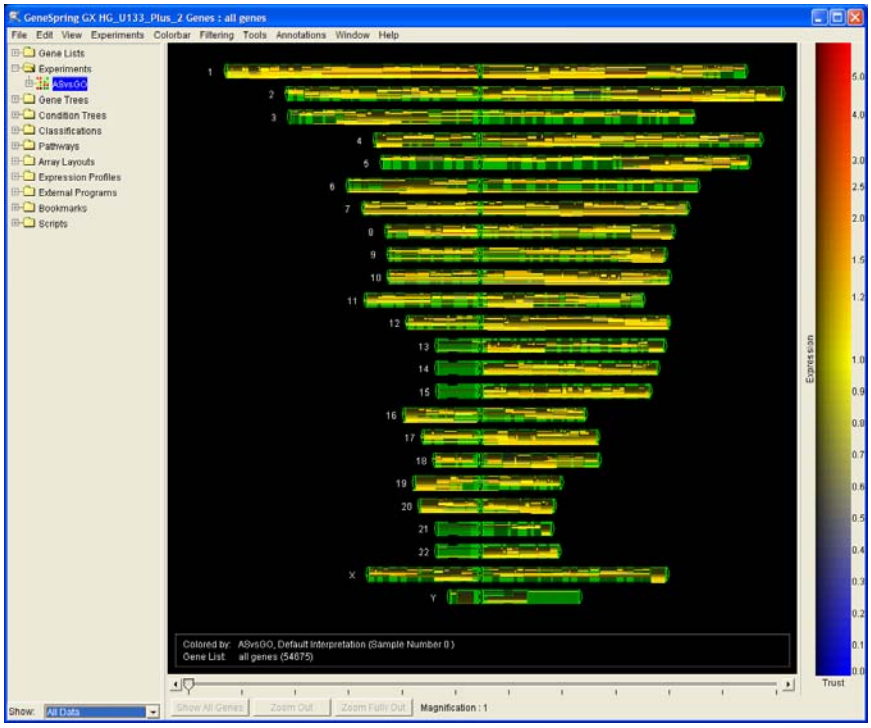
6.4.1.6 Interactive representation of data in genomic context

Visualizing gene expression as a function of the position of the genes on chromosomes is essential in cancer research (e.g. in relation with identification of deleted/amplified regions).

6.4.1.6.1 Detailed specifications

Given the genomic location (e.g. start and end nucleotide coordinates or cytoband location) of a set of genes and an associated set of values (e.g. gene expression, differential gene expression, significance,…), the user should be able to visualize the expression as a function of the position of the gene on chromosomes. Whole-genome, single-chromosome view and partial chromosome view should be available as options, with possibilities to zoom in and out. Individual genes in the graphical representation should be clickable to obtain information about their properties and annotation. Extensions: Amplicon/deletion detection algorithms should be part of the tool, as well as the display of known translocations.

Example: Genome-wide visualization tool in GeneSpring.



6.4.1.7 Interactive 2D and 3D visualization of data

2D- and 3D-plots are the basics of graphical data processing. Generic 2D and 3D visualization tools should be developed, with specialized (subclass) versions for typical applications in genomic context (e.g. MvA plots, Volcano plots in 2D, PCA visualization in 3D). Existing tools include: MayaVi/VTK, OpenDX, XGobi/Ggobi, Vis5d.

6.4.1.7.1 Detailed specifications

Given a set of coordinates (e.g. M and A value), an identifier (e.g. feature ID) for each point and optionally a category identifier and an extra value (continuous, e.g. p-value) associated to the point, the specifications of the viewing tools are:

For 2D plots:

- Zooming in and out
- Coloring the spots according to the category identifier and/or extra value.
- Assigning a symbol according to the category identifier.

- Selecting a set of points in the figure and dump the identifiers of the represented data.
- Dumping the settings of the figure (e.g. axes boundaries) such that it can be reproduced in a script.
- Double-clicking on a point of the figure to open a window describing the feature in greater detail

For 3D plots:

- As for 2D plots plus,
- Rotating the figure.

6.4.1.8 Browsable visualization of data in DAG

Directed Acyclic Graphs (DAG) are a way to code hierarchically representable information, representation being usually done in a tree-like fashion. One important DAG in the context of the clinical studies is the Gene Ontology (GO) which represents the properties of the genes with increasing detail the deeper one goes in the GO tree. Statistical parameters can be associated with (so-called) GO nodes, e.g. comparing the number of times a GO term was found associated with a list of genes found differentially expressed with the number of times that GO term occurs on a microarray. (Note: In R the GOstat package is able to handle the statistical aspects of the problem for the GO.) The developed tool should not be limited to the Gene Ontology, but allow for any DAG-representable data structure.

6.4.1.8.1 Detailed specifications

Given a DAG, a set and a subset of data represented on the DAG.

Basic actions:

- Being able to represent DAGs in a tree-like fashion
- Being able to compute the significance of the nodes in the DAG when comparing a given dataset against a reference, both annotated with the DAG.
- Selecting nodes and dump the associated features in a list.

6.4.1.9 Venn diagram tool

Venn diagrams are a very popular tools to identify genes common to two or three categories.

6.4.1.9.1 Detailed specifications

Given two or three lists of identifiers (and the corresponding list labels).

Basic actions:

- Build a Venn diagram.
- Produce list of identifiers in all regions of the graph.

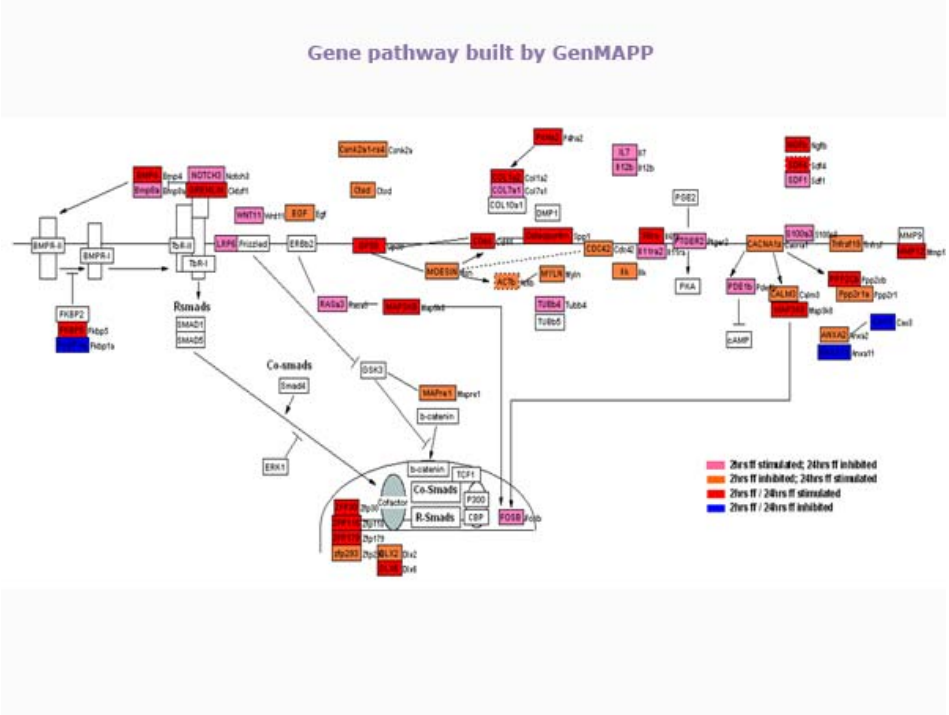
6.4.1.10 Pathway visualization

Interaction pathways are the core of the molecular biology. Databases of known pathways exist, as well as lists of one-to-one relationships between genes (e.g. through automatic literature text mining). The user of the ACGT environment should be able to visualize his/her

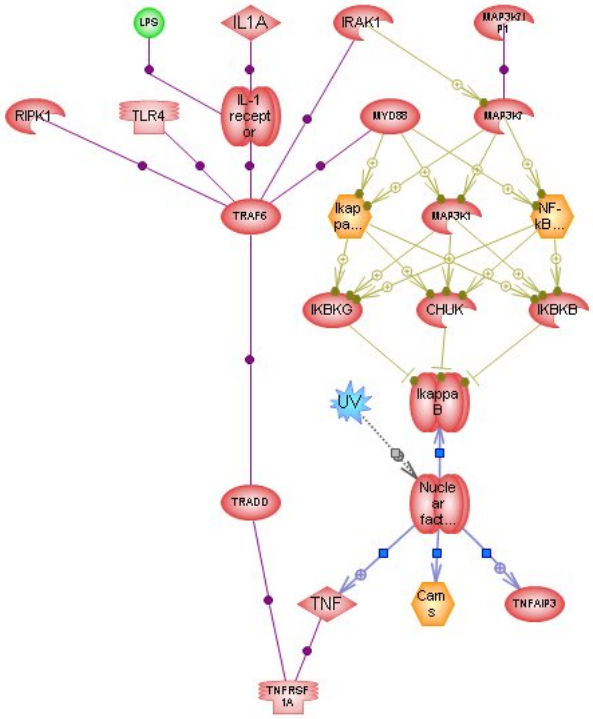
data in the context of these two sources of information. Two basic types of pathway problems exist: (1) representing gene expression in fixed known pathways, and (2) construct interaction pathways based on known (or putative) one-to-one interactions. A source of pathway information maybe Biovista's database.

Existing pathway analysis tools are: Ingenuity Pathway Analysis, PathwayStudio.

Example: GenMAPP plot, based on existing pathway knowledge



Example: PathwayAssist plot (PA is the ancestor of PathwayStudio), based on one-to-one interactions.



6.4.1.10.1 Detailed specifications

Problem 1: Representing gene expression in the context of known pathways

Input information: list of genes and associated parameter (e.g. expression or significance), pathway databases

Basic actions:

- Identify the pathways with the biggest chance to be actually involved from the list.
- Let the user view the pathway with the genes in the list emphasized w.r. to the others in the pathway (use a color scale to represent the associated values).

Problem 2: construct pathway based on one-to-one interactions.

Input information: list of genes, pathway databases

Basic actions:

- Construct an interaction pathway from the set of one-to-one gene interaction database. Different threshold can be used: based on the quality of the one-to-one interactions, based on the maximum distance between genes in the pathway, etc…

Features common to both problems:

- Double-clicking on the node of a pathway opens a window with detailed information about the gene.

- Producing a list of gene identifiers by selecting a set of nodes in the pathways (or dump the entire set of genes in the pathway in a single list).
- User-definable position of the genes in the graph. (This information should be storable such that the graph can be reproduced later.)

6.5 Conclusions

The presented requirements' analysis is not adequate for detailed functional specification of system components, since additional technical and functional specification activities are required for this. On the other hand, adequate requirements have been presented and analysed so that the Technical WPs of the Project begin their work.

This was infact foreseen in the ACGT DoW, in which functional specifications of application required, data access services and analytical services is to take place within the technical Workpackages of the Project. The presented future scenarios to be supported by the ACGT technologies and the requirements elicited provide adequate material for all these WPs to begin their detailed functional specifications and subsequently development of the required applications and services.

Obviously continuous interaction between the various "Scenario Owners" and the technical WPs will be required (See section "Volatility of Requirements" in Chapter 2) and this process is to be continuously supported by the Project Management through appropriate mechanisms (Project collaborative tools, i.e. WiKi, physical meetings and workshops, etc).

PART 2

State of the art review of technological domains relevant to ACGT

7 State of the art review

7.1 Introduction

For the successful implementation of the project a large number of technologies need to be seamlessly integrated. It is worth mentioning that ACGT is not a standard development project but rather a standard adopting one. As a result, it is of particular importance to provide a thorough state-of-the-art in the many domains related to ACGT. On the other hand the project needs to advance state-of-the-art in a number of other scientific domains.

The main areas of relevance to the ACGT workplan are listed below and the corresponding state-of-the-art review done in presented in subsequent chapters of Part 2 of this deliverable. The areas and scientific domains reviewed, following a generic SoA review of current integrated clinico-genomic studies for disease-specific diagnosis, prognosis and therapy design, are:

1. Grid Computing, with emphasis on Grid Computing in the domain of biomedicine.
2. Tools and techniques for the visual invocation and orchestration of services
3. Systems and standards for Clinical Trial Management Systems
4. Tools for the creation and management of clinical trials
5. Tools and techniques for the analysis of biomedical data
6. Approaches to the Integration of heterogeneous Databases
7. Biomedical Ontologies, Terminologies and databases relevant to Oncology
8. In-Silico Modelling and Simulation of tumour response to therapy
9. Data Minining and Knowledge Discovery
10. Metadata Standards
11. Workflow Management and enactment systems
12. Interactive 3D Visualization

In addition to these and based on the ethical and legal constraints of the domain which is presented in Chapter 21, a robust security architecture needs to be defined. The related state of the art is presented in Chapter 22.

The section concludes with a state of the art review related to Online training standards and platforms and Grid Portals, which are also an essential part of the ACGT workplan and envisioned environment.

8 Challenges in integrated "-omic" studies

8.1 Introduction

An objective of the new biomedical informatics technology is to clarify the sensitivity and specificity of genomic medicine. In this context, exploratory analyses is the process of generating hypotheses that are later supported (or not) by the data (e.g. hypothesis: gene x is responsible for a side effect of drug y). The task of validating these hypotheses is done by means of clinical trials. Many clinical trials have problems because they cannot gather enough information to draw sound conclusions in a timely manner – this applies not only to the number of patient subjects but to the lack of links between clinical and genomic patients' data.

We can safely assume that this problem will be aggravated by looking at smaller and smaller groups of patients defined by many genetic characteristics and for this reason it is necessary to provide solutions such as seamless access to much larger databases so that citizens/patients with the appropriate genetic features will be more readily identified. New clinical trials must be carefully designed to include genomic and proteomic data. High-quality biomedical databases are urgently needed to provide a sound scientific basis for diagnostic, treatment stratification and predictive tests.

ACGT will contribute to the alleviation of these problems by allowing for discoveries in the laboratory to be quickly transferred to the clinical management and treatment of patients and obtaining societal benefits. In former times, the discovery of diseases such as tuberculosis or diabetes did not lead to immediate therapies. In some cases, an elapse of more than 60 years was needed to improve therapeutics. New technologies such as in-silico experimentation, Grid or data and text mining are contributing to reduce these periods of time.

In the following section we provide a SoA review in the domain of integrated '-omic' studies for disease-specific diagnosis, prognosis and therapy design and the existing challenges. The objective been, for the readers of this document, to understand the mainy challenges ahead and the relevance of all of the technological domains reviewed thereafter, in truly integrated studies, i.e. studies requiring access, integration, analysis and presentation of multilevel biomedical data. Several cases and international studies are described presenting working hypothesis that the joint analysis of genomic and proteomic data will provide more information for modelling disease susceptibility than either analysis alone.

8.2 The role of integrated "-omic" studies in disease prognosis and diagnosis

"Genome sequencing along with other advances, such as the development of DNA microarrays [FOD1997; SCH1996; BRO1999], which allow the simultaneous measurement of the expression of every single gene in a cellular genome, and high-throughput mass spectral analysis of proteins [LOP2003; GEV2003; MAN2003; WAN2003; JON2003], metabolites [ROE2000; FIE2000; ROE2001; TAY2002] and isotopic tracer distribution

patterns [CHR1999; FIS2003; WIT2001; KLA2003], have challenged the conventional paradigm of biological research.

Rather than examining a small number of genes and/or reactions at any one time, we can now begin to look at gene expression and protein activity in the context of networks and systems of interacting genes and gene products. Because our knowledge of this domain is still largely rudimentary, investigations are now routinely moving from being "hypothesis-driven" to being "data-driven" with analysis based on a search for biologically relevant patterns. These technological advances have created enormous opportunities for accelerating the pace of science. One can now envision the possibility of obtaining a comprehensive picture of the mechanisms underlying the cellular function, its regulation, and the interactions of an organism with its environment" [KLA2003].

In the area of disease prognosis and diagnosis, it becomes apparent that modern high-throughput techniques could assist in the elucidation of single or rather groups/profiles of sensitive molecular (bio)markers of a particular disease contributing greatly in the development of early, disease-specific, accurate diagnosis, but even further to the development of appropriate drugs or other therapeutic treatments. Among the longer-term goals is the reconstruction of the biological interaction networks underlying a particular complex disease and its evolution, which will provide eventually the basis of in silico models of cellular function enabling the doctor to test a particular drug and/or therapeutic treatment and identify the right dose and/or combination of therapeutic strategies personalized for each patient and his/her particular medical history and body function/model.

"While the greatest attention has been to-date paid to gene sequence and transcriptional expression analysis using microarrays, it is becoming increasingly clear that these alone cannot be used to accurately determine cellular function. Rather, a comprehensive analysis of biological systems requires the integration of all fingerprints of cellular function: genome sequence, maps of gene expression, protein expression, metabolic output, and in vivo enzymatic expression (activity). While each of these has significant value on its own, the picture that emerges from any single approach is quite limited in nature. Gene transcription is a necessary but not sufficient condition for high in vivo protein expression. Regulation of translation, RNA and protein stability, and post-translational modifications can alter the linear relationship between message and the corresponding protein [SER2003; ROS2003; REH2003]. Additionally, a protein could be present in high concentration, but it may lack the requisite conditions (substrate concentration, cofactors, etc.) for activity in the actual cellular environment [FEL1997; STE1998b]. Integration of all of these profiles for a systematically perturbed cellular system can provide insight about the function of unknown genes, the relationship between gene and metabolic regulation and even the reconstruction of the gene regulation network [IDE2001; KLA2003].

This need for integration is to some extent clear in the case of complex, multifactorial diseases/ traits, such as obesity, diabetes, hypertension, schizophrenia (and other diseases of the nervous system, including Parkinson's and Alzheimer's) and cancer. Cancer in general and breast cancer in particular is a highly complex and heterogeneous disease which involves a succession of genetic changes that eventually results in the conversion of normal cells into cancerous ones. Hanahan and Weinberg [HAN2000] have proposed that human tumours are governed by a common set of six acquired capabilities;

1) self-sufficiency in growth signals;
2) insensitivity to anti-growth signals;
3) evasion of apoptosis;
4) limitless replicating potential;

5) sustained angiogenesis; and
6) tissue invasion and metastasis.

It is obvious that a complete knowledge of these processes requires the integration and analysis of massive amounts of data as is being collected from current genomic, proteomic and metabolomic platforms [HOO2004]. But it is not only the multiplicity of the factors (and cellular levels) contributing to a particular disease framework that imposes approaching the problem in a systematic way. Even for Mendelian genetic disorders, nearly all of which have now been correlated with a specific gene or set of genes [HOH2004] due to remarkable advances in gene mapping and bioinformatics, the relationship between genotype and phenotype is not as simple as expected (and/or currently treated) [SCR1999; LAI2004]. One can assume – and this concept has dominated bioinformatics analysis practices to-date – that a typical corollary of the central dogma of biology is that phenotype is determined solely by the action of proteins. Adhering to such a model could lead to the hypothesis that measurement of protein levels, and in lack of such large-scale ability to-date, of gene expression levels, alone would be perfectly predictive of a disease. This approach, however, disregards the fact that the different levels of cellular function are also governed by "local" control mechanisms that disrupt the linear relationship between gene expression, protein concentration and reaction rate, while they can affect each other along the "classical" direction of the central dogma, but in a feedback manner as well. These "deviations" from the linear scheme and single direction of the central dogma have been very well documented [MAT2003]. Simply, environmental influence initiates a cascade of signalling reactions (main subject of proteomic research), ending at the activation of certain transcription factors, i.e. proteins, which regulate the expression of the genes. Members of the long interspersed Element-1 (Line-1) family encode the necessary products to ensure retrotranscription. Small interfering RNAs mediate post-transcriptional gene silencing via the RNA interference pathway. Proteins regulate other proteins via ubiquitin functioning in degradation. Outside environmental influences affect the distribution of metabolic fluxes/reaction rates, which in sequence affects the expression of certain genes. It becomes clear then that revealing the mechanisms underlying complex diseases/cellular mechanisms requires combined information from all levels of cellular function. It is evident that measuring gene expression and/or protein concentration and/or metabolite concentration alone might miss vital information reflecting the complexity of biological systems.

Despite the importance of integrated genomic, proteomic and metabolic studies, very few experiments have been done to date that actually combine information from multiple cellular profiles. More recent work has focused on one analysis of a single data type [NEV2003], or at best a combination of genomic and proteomic profiles [IDE2001].

It needs to be underlined that in the case of disease diagnosis, there is no published integrated study to-date. "*One of the main reasons is that we presently lack both the conceptual understanding and the computational tools that would allow the identification of cause-effect relationships between the gene and protein expression and phenotypic profiles. The development, however, of algorithms to address these questions cannot be accomplished in the absence of experimental data that monitor the cellular physiology under a variety of conditions at all stages of growth and levels of cellular function. Taking into consideration the different time-scales of the various biological processes, it is therefore very important to carefully design experiments that can provide comparable gene expression, protein production and metabolic function data that can lead to useful results. This will be closely tied to technological developments aiming at increasing and improving the experimental techniques and computational methodologies for the quantitative measurement of the cellular physiological state at each level of cellular function*" [KLA2003].

Tight to the last sentence is the fact that many challenges in quantitative high-throughput analyses at each individual cellular level still exist [KAN2005]. Therefore, before any integrated analyses could be attempted in a serious way, these challenges need to be resolved. They range from limitations in the available experimental protocols, to lack of data analysis techniques for upgrading the information content of the acquired data. In addition, "holistic analyses of biological systems require a change in the way in which questions are approached in the biological sciences. Collecting, managing, and analyzing comparable data from various cellular profiles requires expertise from several fields that transcend traditional discipline boundaries, including engineering and computer science, statistics and applied mathematics, and chemistry, physics, and biology" [KLA2003]. This might be the most significant challenge, but the also the greatest opportunity for raising the next generation of systems biology researchers who will be able to work, interact and produce in a highly interdisciplinary environment which will certainly lead to novel discoveries and inventions.

In the next three sections, we present in brief the current state-of-the-art in the fields of genomics/transcriptomics, proteomics and metabolomics focusing, if studies have been reported, in disease diagnosis/prognosis in general and in cancer research, in particular. In the final paragraph, we discuss the directions and the current and in the near future anticipated technological, computational and analytical challenges associated with the integrated –omic studies.

8.3 Genomics/Transcriptomics and Disease Prognosis/Diagnosis

New genetic and genomic tools have revolutionized the way in which multifactorial diseases are investigated. Genotyping of vast numbers of genetic polymorphisms in large populations is very important for the identification of the etiologically relevant mutations. Many of these mutations have been discovered as a direct result of recent monumental advances in high-throughput genome screening techniques. These techniques use the recently completed human genome sequence to construct array-based platforms, including expression arrays, exon arrays, sequencing arrays and single nucleotide polymorphism (SNP) arrays [DOB2003].

However, once disease-associated genes have been identified, functional validation approaches are crucial to confirm the role of the candidate genes in diseases process. On the other hand, expression profiling can be an effective approach for developing disease diagnostics, predicting the prognosis of a particular, multifactorial disease and for the identification of abnormally regulated genes and signalling pathways that contribute to cellular dysfunction and disease and therefore they are expected to result in the identification of strong candidate targets for diagnosis and therapeutic intervention [REI2006]. Thus far, whole genome expression analysis has been successfully applied primarily to numerous forms of cancer [DYB2004; NAM2004; PET2004; LI2004; WOO2004; MAZ2004], including breast cancer [YAN2006; for review see NAG2006].

8.4 Proteomics and Disease Prognosis/Diagnosis

It is clear that regulation of translation, protein and RNA stability and posttranslational modifications can affect and alter the linear relationship between message and the corresponding protein. Proteomics refers to systematic study of every protein and protein modification produced by the cell. Proteomic technologies include those that require "a priori" hypothesis and are not based on mass spectrometry such as tissue, protein and antibody

microarrays and those approaches that take advantage of the ability of mass spectrometry to separate peptides and proteins according to their mass-to-charge-ratio. Usually mass spectrometry-based proteomics include an initial phase of protein separation commonly performing 2-Dimensional gel electrophoresis (2D) or currently performing differential in-gel electrophoresis (DIGE).

After successful separation of proteins and subsequent isolation and digestion further forms of proteomic technologies are used. These are 2 dimensional gel electrophoresis – matrix-assisted laser desorption /ionization time-of-flight mass spectrometry (2D-MALDI-TOF-MS) and alternatives to TOF as Q-TOF and TOF/TOF. 2D gel -electrospray ionization tandem mass spectrometry (2D-ESI-MS/MS) uses electrospray ionization (ESI). Alternatively MALDI 2D-LiquidChromatography-Mass Spectrometry (2D-LC-MS or alternatively to MS tandem mass spectrometry (2D-LC-MS/MS)) can be used. Recently TIGE began to substitute 2D in the above technologies in a wide range of studies since it improves speed, sensitivity and reproducibility of the obtained data [for review see BER06].

Also, other mass spectrometry-based and non gel-based proteomic approaches have been developed and used in several tumour samples with good results in terms of diagnostic or prognostic classification [YAN2003; CHA2004; SCH2005; PAW2006] such as Isotope-Coded Affinity Tagging (ICAT) and Multidimensional Protein Identification Technology (MudPIT) coupled with MS/MS or MALDI/MS while alternatively to MS/MS and MALDI/MS, surface enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF-MS) has been used to profile clinical biological fluids such as serum, saliva, and cerebrospinal fluid to identify protein profiles that are diagnostic or prognostic indicators of disease course [PET2002; KOZ2003; WON2003; HAN2003; KOO2004; ZHA2004]

The development of proteomic technologies is still in its infancy and their effectiveness is limited because of the complexity of sequences present and the vast number of possible posttranslational modifications. On average, each mammalian transcript encodes ten different protein isoforms which might be differentially regulated at the protein rather than the mRNA level, greatly increasing, consequently, the complexity of genome and the potential information available. Thus far, cancer has served as the model disease for the application of proteomics to human disease [for a review see BER2006]. Since, in most cancer forms, the affected tissue is clonally derived and relatively homogenous with respect to the cellular phenotype, proteomic technologies have resulted in the identification of novel protein biomarkers that will be useful for early diagnosis, prediction of outcome and monitoring of disease course for multiple forms of cancer [RAI2004; ROD2004].

Moreover, proteomic approaches can be greatly advantageous on biofluids since these tissues are more easily available than affected tissues and can be screened rapidly to identify protein profiles indicative or either predictive of currently developing or ongoing diseases processes or even for the therapeutic follow-up of patients [PET2002; LI2002; PUZ2004; HEI2005; HU2005; SHI2006]. However, once a protein is present in normal levels, the lack of the precise requisite conditions (substrate concentration, cofactors, etc.) can alter its activity in the actual cellular environment.

8.5 Metabolomics/Metabolic Flux Analysis/Pharmacokinetics

Metabolomic profiling is the – most recently introduced [FIE2000; ROE2000], but one of the currently fastest growing- high-throughput platforms for the analysis of the cellular metabolic fingerprint. It refers to the simultaneous quantification of the (relative) concentration of the free small metabolite pools of a biological system [KAN2005] and provides a phenotypic

correspondent of the high-throughput transcriptional and proteomic profiles [FIE2000]. In comparison to genomics and proteomics one advantage of metabolomics is that the human metabolome is estimated to be relatively small (approximately 2,500 unique molecules) [RYA2004] in comparison with the 30,000 genes in the human genome (which are then multiplied through alternate splicing and promoters) as assessed in gene array profiling.

Taking into consideration that changes/differences in the in vivo enzymatic activity between biological systems due to their different responses to certain perturbations and/or their different function are expected to be directly reflected in their metabolomic profiles, the importance of metabolomic profiling analysis becomes quite apparent. “As it is highly unlikely that we will ever be able to develop an in vivo enzymatic activity chip, mapping the flux distribution through a metabolic reaction network [STE1998] is the closest phenotypic equivalent to the type of data we can measure from available techniques for gene and protein expression. Fluxes are determined indirectly from the measurement of net excretion rates of extracellular metabolites and/or the use of isotopically labelled substrates [STE1998]. However, all comprehensive methods for the analysis of complex metabolic flux networks are presently primarily based on steady-state or pseudo-steady-state assumptions in lack of accurate and extensive quantitative measurements of the intracellular metabolite concentrations and their isotopic tracer distribution. Metabolomic profiling is still at its development and standardization stage regarding data validation, normalization and analysis [KAN2005]. Advances in metabolic profiling [FIE2000; ROE2000, ROE2001] defined as the qualitative and quantitative detection (by Nuclear Magnetic Resonance Spectroscopy or Mass Spectrometry) of low-molecular-weight metabolites from the breakdown of the cellular macromolecules, are expected to enhance our understanding of metabolite activity under transient conditions [FIE2003; HAN2003]. This will lead to an increased number of integrated genomic and metabolic studies, which have been currently limited by the lack of flux analysis methodologies for transient physiological conditions. Furthermore, technological and computational developments for metabolic characterization at the microscale [JOH2003] will increase dramatically the number and type of examined physiological conditions opening enormous opportunities in the area of comparative biological studies.” [KLA2003].

Metabolomic profiling was initiated in plant research, where it has been widely used, while only recently it has been applied in bacterial [e.g. KOE2006] and mammalian systems [e.g. UDD2005]. In many biological systems, to which metabolomics has not been widely or even at all used to-date, the questions still concern the initial steps of the experimental protocol, i.e. raw material quantity, selection, acquisition or potential fixation/treatment. Up today, metabolomics is typically performed in biofluids and can be useful for physiological evaluation, drug safety assessment, drug therapy monitoring and diagnosis of human diseases [LIN2004]. Given the novelty of this technology applications to cancer disease have not yet begging.

8.6 Holistic “omic” studies: Current Challenges and Directions

It is obvious that efficient use of the large quantity of data generated from systems biology/integrated “omic” studies requires *development of extended databases that can effectively capture and integrate genomic proteomic and metabolomic data* [KLA2003]. Currently, there exist databases that store DNA and protein-sequence data, protein three-dimensional structure and metabolic pathway structure and stoichiometry, but it is still quite challenging to link information across these diverse resources. Furthermore, these databases should be expanded to accommodate gene and protein expression along with *in vivo* metabolic activity data representing the different phenotypes of multifactorial disorders as cancer. The notion that integration of multiple data types is the only way to truly represent

a complex system flows naturally from the complexity revealed as biologists gain a deeper understanding of common disease aetiologies. A simple model of integrating different data sets from genomic and proteomic data has been developed by Reif and coworkers [REI2004]. This study presents for the first time a working hypothesis that the joint analysis of genomic and proteomic data will provide more information for modelling disease susceptibility than either analysis alone. In the context of simulations performed in this study, it is concluded that the availability of multiple type of data is beneficial when the understanding etiological cause-mechanisms of a particular disease are complex and one or more of the functional variables are missing.

Moreover, there is *a clear need for the development of data visualization and mining software that can be used with diverse data types to explore the relationships that exist and to infer the presence of biological reaction networks/mechanisms* (KLA2003]. Such a system would integrate gene annotation and a variety of expression data to allow the visualization of (metabolic) reaction pathway activity at the transcriptional level, connecting each gene to the reactions that are catalyzed by the protein it encodes. “If one assumes a direct correlation between changes in gene expression and associated enzymatic activity reflected by metabolic output (an assumption in obvious need of verification), gene expression data should allow the formulation of a tentative metabolic (or bio reaction in general) network to be further confirmed by the *in vivo* (metabolic) reaction pathway activity as it is measured in terms of metabolic fluxes or intracellular metabolite or protein concentrations. Any observed inconsistencies, such as high levels of gene expression without a corresponding change in metabolic activity, or the converse, will provide powerful leads to assist in developing verified causal relationships of consequence to overall cell behaviour” [KLA2003]. The contribution of such combined data visualization, integration and combined analysis frameworks in disease prognosis/diagnosis through the identification of biomarkers and reconstruction of networks of cellular mechanisms and correlations of parallel-occurring phenomena, which would have been impossible otherwise, becomes quite obvious. The main objective of ACGT is the development of such a framework in the context of breast cancer research.

Advances and success in predicting disease outcomes based on studies involving existing therapeutic strategies are major steps in the process of bringing systems biology information to clinical practice. It is obvious that we are say goodbye to a past where the patient received the best treatment based solely on historical clinical efficacy data obtained from large populations of patients but without any specific prediction of individual response. We are entering an era where each and every patient will receive individualized therapy based upon the key signalling pathways driving each specific breast cancer form; a systematic integrated platform approach incorporating all “-omics” data sets is critical for moving towards this goal.

8.7 Gene Testing, Pharmacogenomics, and Gene Therapy

DNA underlies almost every aspect of human health, both in function and dysfunction. Obtaining a detailed picture of how genes and other DNA sequences function together and interact with environmental factors ultimately will lead to the discovery of pathways involved in normal processes and in disease pathogenesis. Such knowledge will have a profound impact on the way disorders are diagnosed, treated, and prevented and will bring about revolutionary changes in clinical and public health practice. Some of these transformative developments are described below.

8.7.1 Gene Testing

DNA-based tests are among the first commercial medical applications of the new genetic discoveries. Gene tests can be used to diagnose disease, confirm a diagnosis, provide prognostic information about the course of disease, confirm the existence of a disease in asymptomatic individuals, and, with varying degrees of accuracy, predict the risk of future disease in healthy individuals or their progeny.

Currently, several hundred genetic tests are in clinical use, with many more under development, and their numbers and varieties are expected to increase rapidly over the next decade. Most current tests detect mutations associated with rare genetic disorders that follow Mendelian inheritance patterns. These include myotonic and Duchenne muscular dystrophies, cystic fibrosis, neurofibromatosis type 1, sickle cell anaemia, and Huntington's disease.

Recently, tests have been developed to detect mutations for a handful of more complex conditions such as breast, ovarian, and colon cancers. Although they have limitations, these tests sometimes are used to make risk estimates in presymptomatic individuals with a family history of the disorder. One potential benefit to using these gene tests is that they could provide information to help physicians and patients manage the disease or condition more effectively. Regular colonoscopies for those having mutations associated with colon cancer, for instance, could prevent thousands of deaths each year.

Some scientific limitations are that the tests may not detect every mutation associated with a particular condition (many are as yet undiscovered), and the ones they do detect may present different risks to different people and populations. Another important consideration in gene testing is the lack of effective treatments or preventive measures for many diseases and conditions now being diagnosed or predicted.

Revealing information about the risk of future disease can have significant emotional and psychological effects as well. Moreover, the absence of privacy and legal protections can lead to discrimination in employment and insurance or other misuse of personal genetic information. Additionally, because genetic tests reveal information about individuals and their families, test results can affect family dynamics. Results also can pose risks for population groups if they lead to group stigmatization.

Other issues related to gene tests include their effective introduction into clinical practice, the regulation of laboratory quality assurance, the availability of testing for rare diseases, and the education of healthcare providers and patients about correct interpretation and attendant risks.

Families or individuals who have genetic disorders or are at risk for them often seek help from medical geneticists (an MD specialty) and genetic counsellors (graduate-degree training). These professionals can diagnose and explain disorders, review available options for testing and treatment, and provide emotional support. (For more information, see Medicine and the New Genetics)

8.7.2 Pharmacogenomics: Moving Away from "One-Size-Fits-All" Therapeutics

Within the next decade, researchers will begin to correlate DNA variants with individual responses to medical treatments, identify particular subgroups of patients, and develop drugs customized for those populations. The discipline that blends pharmacology with genomic capabilities is called pharmacogenomics.

More than 100,000 people die each year from adverse responses to medications that may be beneficial to others. Another 2.2 million experience serious reactions, while others fail to respond at all. DNA variants in genes involved in drug metabolism, particularly the cytochrome P450 multigene family, are the focus of much current research in this area. Enzymes encoded by these genes are responsible for metabolizing most drugs used today, including many for treating psychiatric, neurological, and cardiovascular diseases. Enzyme function affects patient responses to both the drug and the dose. Future advances will enable rapid testing to determine the patient's genotype and guide treatment with the most effective drugs, in addition to drastically reducing adverse reactions.

Genomic data and technologies also are expected to make drug development faster, cheaper, and more effective. Most drugs today are based on about 500 molecular targets; genomic knowledge of the genes involved in diseases, disease pathways, and drug-response sites will lead to the discovery of thousands of new targets. New drugs, aimed at specific sites in the body and at particular biochemical events leading to disease, probably will cause fewer side effects than many current medicines. Ideally, the new genomic drugs could be given earlier in the disease process. As knowledge becomes available to select patients most likely to benefit from a potential drug, pharmacogenomics will speed the design of clinical trials to bring the drugs to market sooner.

8.7.3 Gene Therapy, Enhancement

The potential for using genes themselves to treat disease or enhance particular traits has captured the imagination of the public and the biomedical community. This largely experimental field—gene transfer or gene therapy—holds potential for treating or even curing such genetic and acquired diseases as cancers and AIDS by using normal genes to supplement or replace defective genes or bolster a normal function such as immunity.

More than 600 clinical gene-therapy trials involving about 3500 patients were identified worldwide in 2002.[12] The vast majority take place in the United States (81%), followed by Europe (16%). Although most trials focus on various types of cancer, studies also involve other multigenic and monogenic, infectious, and vascular diseases. Most current protocols are aimed at establishing the safety of gene-delivery procedures rather than effectiveness.

Gene transfer still faces many scientific obstacles before it can become a practical approach for treating disease. According to the American Society of Human Genetics' Statement on Gene Therapy, effective progress will be achieved only through continued rigorous research on the most fundamental mechanisms underlying gene delivery and gene expression in animals.

8.8 References

[BER2006] Bertucci, F., et al. (2006) Proteomics of breast cancer: Principles and potential clinical applications. MCP (In press)

[BRO1999] Brown, P.O. and Botstsein, D. (1999) Exploring the new world of the genome with DNA microarrays. Nature Genetics 21: 33-37

[CHA2004] Chaurand, P., et al. (2004) Proteomics in diagnostic pathology: profiling and imaging proteins directly in tissue sections. Am. J. Pathol. 165: 1057-1068

[CHR1999] Christensen, B. and Nielsen, J. (1999) Isotopomer Analysis Using GC-MS. Metab.

[12] Journal of Gene Medicine Web site (www.wiley.co.uk), accessed March 2003

Eng. 1: 282 (E8)-290 (E16)

[DOB2003] Dobrin, S.E. and Stephan, D.A. (2003) Integrated microarrays into disease-gene identification strategies. Expert Rev. Mol. Diagn. 3: 375-385

[DYB2004] Dybkaer, K., et al. (2004) Molecular diagnosis and outcome prediction in diffuse B-cell lymphoma and other subtypes of lymphoma. Clin. Lympoma 5: 19-28

[FEL1997] Fell D. (1997) Understanding the control of metabolism. Portland Press Ltd, London

[FIE2000] Fiehn O., et al. (2000) Metabolite profiling for plant functional genomics. Nature Biotech. 18:1157-1168

[FIE2003] Fiehn, O. and Weckwerth, W. (2003) Deciphering metabolic networks. Eur. J. Biochem. 270: 579-588

[FIS2003] Fischer, E. and Sauer, U. (2003) Metabolic flux profiling of Escherichia coli mutants in central carbon metabolism using GC-MS. Eur. J. Biochem. 270: 880-891

[FOD1997] Fodor, S.P.A. (1997) Massively parallel genomics. Science 277: 393-395

[GEV2003] Gevaert, K, et al., (2003) Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. Nat. Biotechnol. [epub ahead of print]

[HAN2000] Hanahan, D. and Weinberg, R.A. (2000) The hallmarks of cancer. Cell 100: 57-70

[HAN2003a] Hanash, S. (2003) Disease proteomics. Nature 422: 226-232

[HAN2003b] Hans, M.A., et al. (2003) Free intracellular amino acid pools during autonomous oscillations in Saccharomyces cerevisiae. Biothechnol. Bioeng. 82: 143-151

[HEI2005] Heike, Y., et al. (2005) Identification of serum proteins related to adverse effects induced by docetaxel infusion from protein expression profiles of serum using SELDI ProteinChip system. Anticancer Res. 25: 1197-1203

[HOH2004] Hoh, J. and Ott, J (2004) Genetic dissection of diseases: design and methods. Curr. Opin. Gen. Dev. 14: 229-232

[HOO2004] Hood, L., et al. (2004) Systems biology and new technologies enable predictive and preventative medicine. Science 306: 640-643

[HU2005] Hu, Y., et al. (2005) SELDI-TOF-MS: the proteomics and bioinformatics approaches in the diagnosis of breast cancer. Breast 14: 250-255

[IDE2001] Ideker, T., et al. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292: 929-934

[JOH2003] John, G.T., et al., (2003) Integrated optical sensing of dissolved oxygen in microtiter plates: A novel tool for microbial cultivation. Biothechnol. Bioeng. 81: 829-836

[JON2003] Jones, J.J., et al. (2003) Investigation of MALDI-TOF and FT-MS techniques for analysis of Escherichia coli whole cells. Anal Chem. 75:1340-7

[KAN2005] Kanani, H. and Klapa, M.I. (2005) Data Correction Strategy for Metabolomics Analysis using Gas Chromatography-Mass Spectrometry (under review in Metabolic Engineering)

[KLA2003] Klapa, M. and Quachenbush, J. (2003) The quest of the mechanisms of life. Biothech. & Bioeng. 84: 739-742

[KOE2006] Koek, M.M., et al. (2006) Microbial metabolomics with gas chromatography / mass spectopmetry. Anal. Chem. 78, 1272-1281

[KOO2004] Koopmann, J., et al. (2004) Serum diagnosis of pancreatic adenocarcinoma using surface-enhanced laser desorption and ionization mass spectrometry. Clin. Cancer Res. 10: 860-868

[KOZ2003] Kozak, K.R., et al. (2003) Identification of biomarkers for ovarian cancer using strong anion-exchange ProteinChips: Potential use in diagnosis and prognosis. Proc. Natl. Acad. Sci. (USA) 100: 12343-12348

[LAI2004] Lai, K. and Klapa, M. (2004) Alternative pathways of galactose assimilation: could

inverse metabolic engineering provide an alternative to galactosemic patients? Metab. Eng. 6: 239-244

[LI2002] Li, J., et al. (2002) Proteomics and bioinformatics approaches for identification of serum biomar4kers to detect breast cancer. Clin Chem. 48: 1296-1304

[LI2004] Li, S.R., et al. (2004) Differential expression patterns of the insulin-like growth factor 2 gene in human colorectal cancer. Tumour Biol. 25: 62-68

[LIN2004] Lindon, J.C., et al. (2004) Metabolomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis. Biomarkers 9: 1-31

[LOP2003] Lopez F., et al. (2003) Improved sensitivity of biomolecular interaction analysis mass spectrometry for the identification of interacting molecules. Proteomics 3:402-12

[MAN2003] Manabe, T. (2003) Analysis of complex protein-polypeptide systems for proteomic studies. J Chromatogr B Analyt Technol Biomed Life Sci. 787:29-41

[MAT2003] Mattick, J.S. (2003) Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. Bioessays 25: 930-939

[MAZ2004] Mazzanti, C., et al. (2004) Using gene expression profiling to differentiate benign versus malignant thyroid tumors. Cancer Res. 64: 2898-2903

[NAG2006] Nagasaki, K. and Miti, Y. (2006) Gene expression profiling of breast cancer. Breast Canc. 13:2-7

[NAM2004] Nambiar, S., et al. (2004) Applications of array technology: melanoma research and diagnosis. Expert Rev. Mol. Diagn. 4: 549-557

[NEV2003] Nevins, J.R., et al. (2003) Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. Hum. Mol. Cen. 12: R153-R157

[PAW2006] Pawlik, T., et al. (2006) Proteomic analysis of nipple aspirate fluid from women with early-stage breast cancer using isotope-coded affinity tags and tandem mass spectrometry reveals differential expression of vitamin D binding protein. BMC Cancer 6: 68

[PET2002] Petricoin, E.F., et al. (2002) Use of proteomics patterns in serum to identify ovarian cancer. Lancet 14: 878-885

[PET2004] Petty, R.D., et al. (2004) Gene expression profiling in non-small cell lung cancer: from molecular mechanisms to clinical application. Clin. Cancer Res. 10: 3237-3248

[PUS2004] Pusztai, L., et al. (2004) Pharmacoproteomic analysis of prechemotherapy and postchemotherapy plasma samples from patients receiving neo adjuvant or adjuvant chemotherapy for breast cancer. Cancer 100: 1814-1822

[PYA2004] Pyals, J. (2004) Metabolomics – An important emerging science, Drug Discovery Metabolomics. Pharmatech: 51-53

[RAI2004] Rai, A.J. and Chan, D.W. (2004) Cancer proteomics: serum diagnostics for tumor marker discovery. Ann. N.Y.Acad. Sci. 1022: 286-294

[REH2003] Rehfeld, J.F. and Goetze, J.P. (2003) The posttranslational phase of gene expression: new possibilities in molecular diagnosis. Curr Mol Med. 3:25-38

[REI2004] Reif, D.M., et al. (2004) Integrated analysis of genetic, genomic and proteomic data. Expert Rev. Proteomics 1: 67-75

[REI2006] Reis-Filho, J.S., et al. (2006) The impact of expression profiling on prognostic and predictive testing in breast cancer. J. Clin. Pathol. 59: 225-231

[ROD2004] Rodland, K.D. (2004) Proteomics and cancer diagnosis: the potential of mass spectrometry. Clin. Biochem. 37: 579-583

[ROE2000] Roessner, U., et al. (2000) Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. Plant J. 23:131-142

[ROE2001] Roessner, U., et al. (2001) Metabolic Profiling Allows Comprehensive Phenotyping of Genetically or Environmentally Modified Plant Systems. Plant Cell 13:11-29

[ROS2003] Rossignol, F., et al. (2003) Expression of lactate dehydrogenase A and B genes in different tissues of rats adapted to chronic hypobaric hypoxia. J Cell Biochem. 89:67-79

[SCH1996] Schena, M., et al. (1996) Parallel human genome analysis: Microarray based expression monitoring of 1000 genes. Proc. Natl Acad. Sci. (USA) 93: 10614-10619

[SCH2005] Schwartz, S.A., et al. (2005) Proteomic-based prognosis of brain tumor patients using direct-tissue matrix-assisted laser desorption ionization mass spectrometry. Cancer Res. 65: 7674-7681

[SCR1999] Scriver, C.R. and Waters, P.J. (1999) Monogenic traits are not simple: lessons from phenylketonuria. Trends Genet. 15: 267-272

[SER2003] Serikawa, K.A., et al. (2003) The transcriptome and its translation during recovery from cell-cycle arrest in Saccharomyces cerevisiae. Mol Cell Proteomics. [epub ahead of print]

[SHI2006] Shin, B.K., et al. (2006) Proteomic approaches to uncover the repertoire of circulating biomarkers for breast cancer. J. Mam. Gland Biol & Neopl. 7: 407-413

[STE1998a] Stefanopoulos, G. (1998) Metabolic fluxes and metabolic engineering. Metab. Eng. 1: 1-10

[STE1998b] Stephanopoulos G., Aristidou A., Nielsen J. 1998. Metabolic Engineering. Academic Press, San Diego

[TAY2002] Taylor, J., et al. (2002) Application of metabolomics to plant genotype discrimination using statistics and machine learning. Bioinformatics Suppl 2:S241-S248

[UDD2005] Uddin, P.K., et al. (2005) Towards unraveling ethanol-specific neuro-metabolomics based on ethanol responsive genes in vivo. Neurochem. Res. 30, 1179-1190

[WAN2003] Wang, H. and Hanash, S. (2003) Multi-dimensional liquid phase based separations in proteomics. J Chromatogr B Analyt Technol Biomed Life Sci. 787:11-8

[WIT2001] Wittmann, C. and Heinzle, E. (2001) Application of MALDI-TOF MS to lysine-producing Corynebacterium glutamicum - A novel approach for metabolic flux analysis. Eur. J. Biochem. 268, 2441-2455

[WON2003] Won, Y., et al. (2003) Pattern analysis of serum proteome distinguishes renal cell carcinoma from other urologic diseases and healthy persons. Proteomics 3: 2310-23169

[WOO2004] Woo, I.S., et al. (2004) Expression of placental growth factor gene in lung cancer. Tumour Biol. 25:1-6

[YAN2003] Yanagisawa, K., et al. (2003) Proteomic patterns of tumour subsets in non-small-cell lung cancer. Lancet 362: 433-439

[YAN2006] Yano, K., et al. (2006) A new method for gene discovery in large-scale microarray data. Nucl. Acid Res. 34: 1532-1539

[ZHA2004] Zhang, Z., et al. (2004) Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. Cancer Res. 64: 5882-5890

9 Biomedical Grid Computing

The Grid is widely seen as a step beyond the Internet, incorporating pervasive high-bandwidth, high-speed computing, intelligent sensors and large-scale databases into a seamless pool of managed and brokered resources, available to industry, scientists and the man in the street.

The name, Grid, draws an analogy between the pervasive availability of electrical power and that of computing and data coupled with mechanisms for their effective use. The potential benefits and social impact of the Grid are very great. For the scientist, the Grid will enable large-scale scientific collaborations.

Grid systems and applications aim to integrate, virtualise, and manage resources and services within distributed, heterogeneous, dynamic “virtual organizations”. The realization of this goal requires the disintegration of the numerous barriers that normally separate different computing systems within and across organizations, so that computers, application services, data, and other resources can be accessed as and when required, regardless of physical location.

Grid computing enables the virtualization of distributed computing over a network of heterogeneous resources—such as processing, network bandwidth and storage capacity—giving users and applications seamless, on demand access to vast IT capabilities. Leveraging open standards, it enables computing capacity to be dynamically procured and released based on variations in peak loads—offering business value in day-to-day operations as well as disaster response and recovery.

Grid computing provides a novel approach to harnessing distributed resources, including applications, computing platforms or databases and file systems. Applying Grid computing can drive significant benefits by improving information access and responsiveness, and adding flexibility, all crucial components of solving the data warehouse dilemma.

Rather than bringing data to a data warehouse where it sits hoping to be used, a federated solution can maintain the data at its points of origin. Federated solutions help to address the size and complexity of data ware-houses by applying a logical model to the existing physical infrastructure instead of imposing a new data warehouse environment. Information Grid technology—which gives users and applications security-rich access to virtually any information source, anywhere, over any type of network—supports sharing of data for processing and large-scale collaboration. It also helps bring the federated model to distributed and complex data sources.

Grid computing, also, introduces a new concept to IT infrastructures because it supports distributed computing over a network of heterogeneous resources and is enabled by ***open standards.*** As a result, new and innovative approaches are evolving for harnessing the vast and unused computational power of the world's computers and direct it at research designed to help unlock genetic codes that underlie diseases like AIDS and HIV, Alzheimer's and cancer. A good example of this process is the World Community Grid.

Building on both Grid and Web services technologies, the Open Grid Services Infrastructure (OGSI) defines mechanisms for creating, managing, and exchanging information among entities called Grid services. Succinctly, a Grid service is a Web service that conforms to a set of conventions (interfaces and behaviours) that define how a client interacts with a Grid service. These conventions, and other OGSI mechanisms associated with Grid service

creation and discovery, provide for the controlled, fault-resilient, and secure management of the distributed and often long-lived state that is commonly required in advanced distributed applications.

Examples of biomedical Grid-related projects, with which the project intends to liaise and utilise available work, are:

- *Cancer Biomedical Informatics Grid (caBIG):* It is a cancer-based biomedical informatics network developed by NCICB. caBIG aims to connect cancer related data sources, tools, individuals, and organizations, and to help redefine how research is conducted, care is provided, and patients and participants interact with the biomedical research enterprise.

- *Biomedical Informatics Research Network (BIRN):* The project aims at federating neuroimaging data. BIRN is deploying compute-storage clusters at research and clinical sites around the United States and is deploying Grid middleware to enable the integration of neuroimaging data from multiple locations for the proposes of research and, ultimately, improved clinical care.

- *Shared Pathology Informatics Network (SPIN):* The SPIN initiative at NCI develops an Internet-based software infrastructure to support a network of tissue specimen datasets and associated clinical and pathologic data, needed for cancer research.

- *MEDIGRID*: The project is financed by the French Ministry of Research and investigates the application of Grid technologies for manipulating large medical image databases.

- *MyGRID:* It is a UK EPSRC-funded e-Science pilot project that has developed a comprehensive loosely-coupled suite of middleware components specifically to support data intensive in silico experiments in biology. The project focuses on the semantically rich problems of dynamic resource discovery, workflow specification, and distributed query processing, as well as, provenance management, change notification, and personalization.

- *BioMOBY:* This project aims to develop web services architecture for bioinformatics. In phase two of the BioMOBY project, the plan is to combine web services and the semantic web.

- *Chinook:* It is a peer-to-peer bioinformatics platform whose goal is to facilitate the exchange of analysis techniques and resources within a local community and worldwide.

- *DiscoverySpace:* This project provides a graphical user interface for visualization and analysis of DISCOVERYdb, a centralized data warehouse. The software allows the analysis of biological data without the need for the researcher to access multiple data sources and formulate complex Structured Query Language (SQL) queries and scripts. Its primary focus is for interpreting Serial Analysis of Gene Expression (SAGE) data.

The most important of these projects and their achievements todate are presented in the following sub-section.

9.1 International Projects and Initiatives relevant to ACGT

A large number of Grid related projects have produced results that may prove useful in the context of ACGT. See http://www.Gridstart.org/index.shtml for a detailed list of Grid related projects.

Of specific importance and relevance to ACGT are specific Grid related projects in the domain of Biomedical Informatics Grid and e-science Grid. Below is a short presentation of such projects and their key contributions and results to date – which will be fully explored and utilised by the ACGT project.

The description of the caBIG, the myGrid and BRIDGES projects, their objectives, architectrure and services developed, which are presented in the following sub-sections was made by using published scientific papers by these projects and by using the available information in their respective websites.

The reason for their inclusion in this document is the fact that there is a lot to be learned by the ACGT designers and developers from the approaches taken by these projects, the problems identified and the alternative implementations of various critical services.

9.1.1 caBIG

To expedite the cancer research communities' access to key bioinformatics tools, platforms and data, the NCI is working in partnership with the Cancer Center community to deploy an integrating biomedical informatics infrastructure: caBIG™ (cancer Biomedical Informatics Grid™). caBIG™ is creating a common, extensible informatics platform that integrates diverse data types and supports interoperable analytic tools in areas including clinical trials management, tissue banks and pathology, integrative cancer research, architecture, and vocabularies and common data elements [www.caBIG.org].

The current inventory of applications (tools) and documentation, infrastructure and datasets used to support the caBIG™ initiative can be accessed below. This inventory includes key infrastructure and applications from the NCICB, as well as the various participating Cancer Centers. As the caBIG™ project activities continue, additional interoperable tools, infrastructure, and data will be made available on this site.

caGrid [https://cabig.nci.nih.gov/workspaces/Architecture/caGrid/] provides the core enabling infrastructure necessary to compose the Grid of caBIG. It is a service-oriented architecture and provides the implementation of the required core services, toolkits and wizards for the development and deployment of community provided services, APIs for building client applications, and some sample client applications for interacting with the current test bed installation.

9.1.1.1 Data Models and Metadata

caBIG has adopted a model-driven architecture best practice and requires that all data types used on the Grid are formally described and semantically harmonized. These efforts result in the identification of common data elements, controlled vocabularies, and object-based abstractions for all cancer research domains. caGrid leverages existing NCI data modeling infrastructure to manage, curate, and employ these data models. caCORE is NCICB's platform for data management and semantic integration, built using formal techniques from the software engineering and computer science communities. caCORE defines a data model

specified using industry standard techniques to define common biological constructs (objects). The main components of caCORE include:

- ***Cancer Bioinformatics Infrastructure Objects (caBIO):*** platform independent APIs that reflect an object-oriented view of biomedical information.
- ***Cancer Data Standards Repository (caDSR)***: A metadata registry based upon the ISO/IEC11179 standard that is used to register the descriptive information needed to render cancer research data reusable and interoperable. Data types, such as the caBIO data classes and the data elements on NCI-sponsored clinical trials case report forms, are defined in caCORE UML and converted into ISO/IEC 11179 Administered Components, which are in turn registered in the caDSR.
- ***Enterprise Vocabulary Services (EVS)***: Controlled vocabulary resources that support the life sciences domain, implemented in a description logics framework. EVS vocabularies provide the semantic 'raw material' from which data elements, classes and objects are constructed.

9.1.1.2 The caGRID

In caGrid, both the client and service APIs are object oriented, and operate over well-defined and semantically described data types. Clients and services communicate through the Grid using Globus Toolkit's Grid clients and service infrastructure, respectively. The Grid communication protocol is XML, and thus the client and service APIs must transform the transferred objects to and from XML. This XML serialization of caGrid objects is restricted, as each object that travels on the Grid must do so as XML which adheres to an XML schema registered in the Global Model Exchange (GME). GME defines the syntax of the XML serialization of the properties, relationships, and semantics of caBIG data types.

Furthermore, Globus services are defined by the Web Service Description Language (WSDL). The WSDL describes the various operations the service provides to the Grid. The inputs and outputs of the operations, among other things, in WSDL are defined by XML schemas.

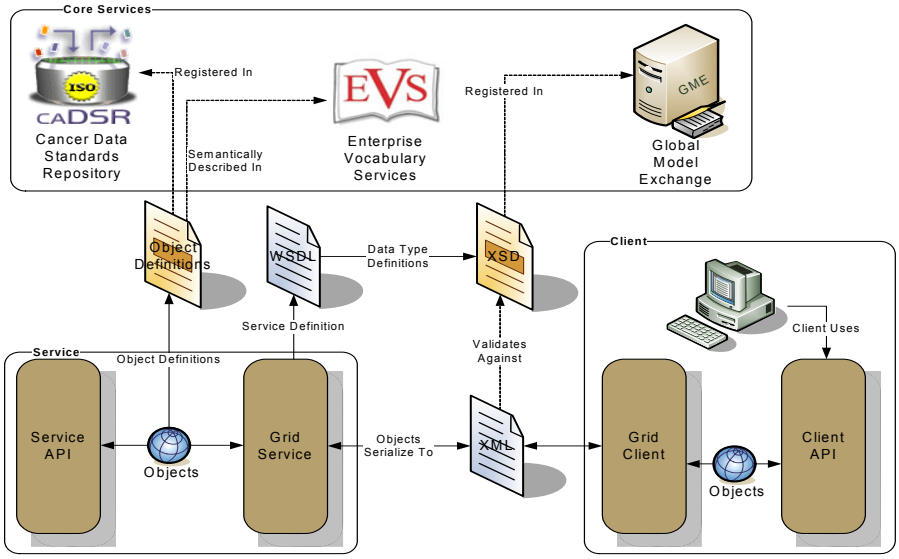


Figure 10: The Grid infrastructure (caGrid) in caBIG

As caBIG requires that the inputs and outputs of service operations use only registered objects, these input and output data types are defined by the XML schemas (XSDs) which are also registered in GME. In this way, the XSDs are used both to describe the contract of the service and to validate the XML serialization of the objects which it uses. The figure above details the various services and artifacts related to the description of and process for the transfer of data objects between clients and services.

9.1.1.3 Services

In the caGrid architecture there are two kinds of services supported:

- Data services that provide access to the caBIG data bases and other data sources
- Analytical services that are used for the processing, transformation, analysis, etc of data

The data services present an object oriented view of data sources and currently support only query functionality. The objects exposed by the data services comply with the common data elements that are registered in the caDSR using the EVS vocabularies.

Their implementation is based on the Open Grid Services Architecture Data Access Integration (OGSA-DAI - www.ogsadai.org.uk) framework which provides a set of interfaces and runtime support for implementing and deploying data sources as Grid services. In order to support the caDSR/EVS semantic integration that is needed when accessing a data source the OGSA-DAI framework has been extended to support a custom XML based query language. A caGrid specific query is formed based on the information extracted from caDSR and submitted, in the OGSA-DAI compliant way, as an Activity in a Perform document, to a data source. The results of the query can be transformed and delivered in any way supported by the standard OGSA-DAI framework.

Analytical services take as input and return strongly typed and semantically harmonized data types as well. These services are necessary in order to support use cases where merely accessing the data is not enough (e.g. when some processing of data is desirable). Analytical services are presented to the caBIG as Grid Services and their implementation is based on the Globus Toolkit. The Globus Toolkit is a reference implementation of Open Grid Services Architecture (OGSA) and it's the Grid platform that caGrid builds upon. A graphical tool is also available to automatically create source code, configuration files, and build process for new analytical services using the GME client to extract the schemas for the types which are required by the analytical service interface.

9.1.1.4 Semantic Discovery in caBIG

A critical requirement of caBIG's infrastructure is that it supports the ability of researchers to discover the available resources. caGrid enables this ability by taking advantage of the rich structural and semantic descriptions of data models and services that are available. The overall architecture for service advertisement and discovery is shown in the figure below.

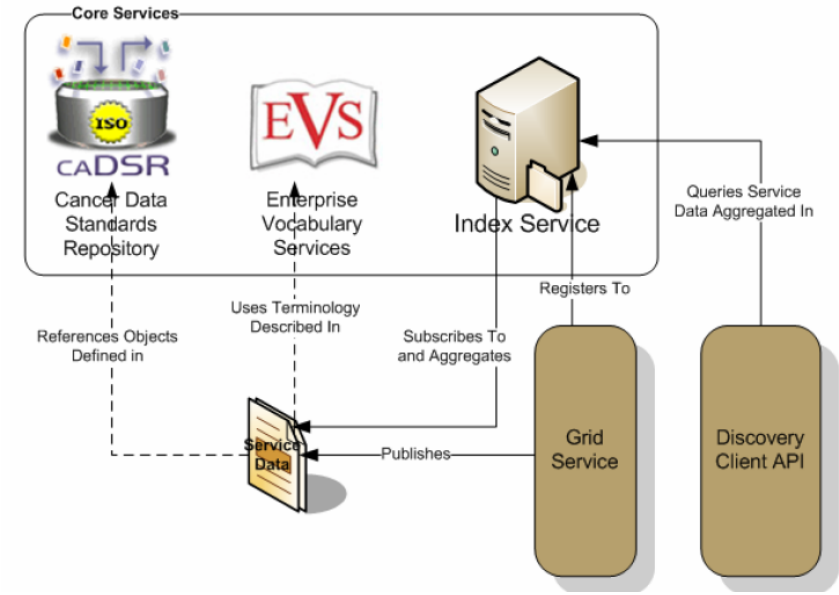


Figure 11: Resource discovery in caBIG

Each service is required to describe itself using caGrid standard service metadata. When a Grid service is connected to the caBIG Grid, it registers its availability and service metadata with a central indexing registry service (Globus Toolkit's Index Service). This service can be thought of as the "yellow pages" and "white pages" of caBIG. A researcher can then discover services of interest by looking them up in this registry using high-level APIs and user applications.

caGrid provides standards for service metadata to which all services must adhere. The basic metadata supported is the Common Service Metadata standard that every service in caBIG is required to provide. This metadata contains information about the service-providing cancer center, such as the point of contact and the institution's name. Extending beyond this generic metadata there are two standards that are specialized depending on whether a data or analytical service is described. The Data Service Metadata details the domain model from which the Objects being exposed by the service are drawn. Additionally, the definitions of the Objects themselves are described in terms of their underlying concepts, attributes, and associations to other Objects being exposed. Similarly, the Analytical Service Metadata details the Objects using the same format as the Data Service Metadata. In addition to detailing the Objects definitions, the Analytical Service Metadata defines the operations the service provides. The input parameters and output of the operations are defined by referencing the appropriate Object definition. In this way, both the data and analytical services fully define the domain objects they expose by referencing the data model registered in caDSR, and identify their underlying semantic concepts by referencing the information in EVS.

The caGrid discovery API and tools allow researchers to query the Index Service for services satisfying a query over the service metadata. That is, researchers can lookup services in the registry using any of the information used to describe the services. For instance, all services from a given cancer centre can be located, data services exposing a certain domain model or objects based on a given semantic concept can be discovered, as can analytical services that provide operations that take a given concept as input.

9.1.1.5 Security and User Management

Security is an especially important component of caBIG both for protecting intellectual property and ensuring protection and privacy of patient related and sensitive information. When security is implemented in a multi-institutional environment, such as caBIG, a challenging problem is to facilitate the management of users and user attributes. Furthermore, it is important to be able to leverage existing systems for authenticating and authorizing requests to the corresponding data sources.

In caGrid the following requirements have been identified:

- ***Secure Communication***
  - *Authentication* – Parties involved can be assured of one another identity
  - *Message Integrity* –Message sent by either party is guaranteed to same message when it is received.
  - *Privacy* – Communication between the two parties can only be interpreted by the two parties
- ***Access Control/Authorization***
  - caBIG services should be able to decide, which users or services may access them.
- ***User/Organizational Attribute Management***
  - Services should have a method for determining the attributes of a requesting party. Such attributes may be needed to service the request. For example, a username and password may be needed to perform a query on a relational database on the party's behalf.
- ***Delegation***
  - caBIG services, should be able to interact with other caBIG services on a user's behalf.
- ***Single Sign On***
  - Users and Grid Services should have one method of authenticating themselves to the Grid; all services in the Grid should accept this method.
- ***User / Organization Management***
  - User/Credential creation should be done within each individual organization. Organizations should have a scalable method of creating and revoking credentials. Services running in one organization should be able to be leveraged by users with credentials in other organizations, assuming there is a trust between the two organizations and the user is authorized to use the service.
- **Virtual Organization**
  - Organizations consisting of users from varying institutions should be able to be grouped together based on a common set of attributes; we refer to these groups of users as a *virtual organization.* caGrid will require infrastructure for creating and managing virtual organizations.

The caGrid security architecture is composed of several components. Components can be classified as *core components* and *external components*. Core components are required by the architecture and are essential for meeting the security requirements for caBIG. External components are those that are considered extensions of the core security architecture. The caGrid security architecture is composed of five core components:

1. **Grid Security Infrastructure (GSI):** Globus provided security infrastructure. GSI provides a method of establishing secure communication between Grid services. This is accomplished using public key cryptography and certificates. Users and Grid services authenticate with one another by exchanging certificates. In most cases users don't actually send around their "real certificate", but rather they create a self

signed short term expiring certificate called a *proxy*. A proxy consists of a new certificate (with a new public key in it) and a new private key. The new certificate contains the owner's identity and a time notation after which the proxy should no longer be accepted by others. The time notation is generally short, giving proxies a short lifetime. Since proxies are used to authenticate with Grid services, they can also be used as a method for single sign on. GSI supports delegation of credentials by allowing users to specify that their proxy certificate can be used by the invoked service to create a new proxy (containing the user's identity) such that the invoked service may act on the user's behalf. The caGrid security architecture uses GSI for single-sign-on, delegation, and secure communication.

2. **Authorization Manager:** Authorization callback mechanism. Globus Toolkit provides a very powerful framework for configuring authentication, but has fairly limited authorization capabilities. The Authorization Manager component of the caBIG Security Framework extends Globus's Authorization Framework to enable the necessary authorization capabilities. As later versions of the Globus Toolkit add more support for configuring authorization, the Authorization Manager is designed to work in the current environment, but compatible with later versions.
3. **Grid User Management Service (GUMS):** Grid Service for the management and creation of Grid users and Grid user credentials. GUMS coordinates a registration process where prospective users can apply for a user account by submitting required information as specified by an administrator. GUMS also supplies interfaces for administrators to review applications and approve or reject them. Given administrative approval, GUMS coordinates the creation of an account which includes the creation of Grid credentials. GUMS allows users access to their credentials, such that they may obtain them and use them to invoke Grid services that require them. With GUMS, Grid users no longer need to worry about all the complexities of managing Grid credentials.
4. **caGrid Attribute Management Service (CAMS):** Grid Service for the management of user/virtual organization attributes. CAMS manages attributes by the user's Distinguished Name (DN), which is found in the user's certificate. For each user CAMS build a database of attributes, in which it persists and manages the user's attributes. The attributes managed for each user can be names, addresses, email, phone numbers, etc. CAMS requires attributes to be represented in XML format, with each attribute conforming to an XML schema that is stored in GME. Services running on the Grid can use CAMS for requesting a user's attributes such that they can make dynamic runtime decisions. Such decisions would include authorization decisions or business logic routing decisions. CAMS also provides an access control system for protecting attributes, such that only granted user's and services may access a user's attributes.
5. ***Grid Virtual Organization Service (GVOS)***: Grid Service for the management of virtual organizations. This is not currently implemented (as of version 0.5 of caGrid) and will be part of future releases.

External security components that can be used for enhancing the authorization capabilities of caGrid are local authentication/authorization systems or Grid Authorization Services.

9.1.2 My Grid

The information presented in this section on the myGrid project is taken from its web site and the various publications of project partners listed in the references section.

myGrid is an e-Science research project developing open source high-level middleware to support *in silico* experiments in biology. *In silico* experiments use databases and computational analysis tools rather than laboratory investigations to test hypothesis. This section provides an overview of services the myGrid project is developing, and the architecture in which they fit.

Registries provide information about available data and computational services, while remote legacy bioinformatics applications are wrapped using a consistent distributed analysis framework Soaplab. As in conventional science, experimental method is as important as final results. myGrid formalises these methods as workflow or query specifications and provides service based middleware components to enact them.

Personalisation forms a key theme in myGrid service design. Information repositories, service registries and change notification systems are all being developed to provide personalised views of resources. myGrid components make extensive use of metadata to support this need for personalisation and the project is pioneering the use of semantic web technology, to manage annotation, ontologies and semantic discovery. The ultimate goal of myGrid is to supply this collection of services as a toolkit to build end applications. To demonstrate this concept the project is building its own application (the myGrid workBench).

myGrid develops open source high-level service-based middleware to support in silico experiments in biology. In silico experiments are procedures using computer based information repositories and computational analysis adopted for testing hypothesis or to demonstrate known facts. In our case the emphasis is on data intensive experiments that combine use of applications and database queries. The user is helped to create workflows (a.k.a. experiments), sharing and discovering others' workflows and interacting with the workflows as they run. Rather than thinking in terms of data Grids or computational Grids we think in terms of Service Grids, where the primary services support routine in silico experiments. The project's goal is to provide middleware services as a toolkit to be adopted and used in a "pick and mix" way by bioinformaticians, tool builders and service providers who in turn produce the end applications for biologists.

The target environment is open, implying that services and their users are decoupled. Services are not just used solely by their publishers but by users unknown to the service provider, who may use them in unexpected ways. myGrid focuses on speculative explorations by a scientist to form discovery experiments. These evolve with the scientist's thinking, and are composed incrementally as the scientist designs and prototypes the experiment. Intermediate versions and intermediate data are kept, notes and thoughts are recorded, and parts of the experiment and other experiments are linked together to from a network of evidence, as we see in bench laboratory books.

Discovery experiments by their nature presume that the e-biologist is actively interacting with and steering the experimentation process, as well as interacting with colleagues (in the simplest case by email). [GOB2003] gives a detailed motivation for the project.

The project produced a requirements gathering prototype based on use cases for the functional analysis of clusters of proteins. Data was based on microarray studies, which showed the level of activity of genes associated with circadian rhythms in the fruit fly, Drosophila melanogaster They then developed a more detailed set of scenarios for the examination of the genetics of Graves' disease, an immune disorder causing hyperthyroidism [STEa]. This latter case study wa the test bed application for the initial full prototype and the rest of the project. The project has built an electronic laboratory workbench demonstrator application as a vehicle to crystallise our architecture and experiment with our services: their

functionality, their deployment and their interactions [STE2003]. In addition, Talisman is a third party application that is prototyping the use of the various workflow components.

The remaining of the subsections are organised as follows. Section 2 gives an overview of the services in myGrid, its chief components and its architecture. Section 3 runs through an example of the workbench demonstrator. We conclude in section 4 with a statement on status and an outlook for the remainder of the project.

9.1.2.1 myGrid Services and Architecture

The myGrid middleware framework employs a service-based architecture, firstly prototyped with Web Services but with an anticipated migration path to the Open Grid Services Architecture (OGSA) [TAL2002];

[PARWAT] gives an account of the conversion of two myGrid services to OGSI services. The middleware services are intended to be collectively or selectively adopted by bioinformaticians, tool builders and service providers who in turn produce the end applications for biologists. The primary services to support routine *in silico* experiments fall into four categories:

- services that are the **tools** that will constitute the experiments, that is: specialised services such as AMBIT text extraction [GAIHEP], and external third party services such databases, computational analysis, simulations etc, wrapped as web services by Soaplab if required;

- services for **forming and executing experiments**, that is: workflow management services[ADDOIN], information management services, and distributed database query processing [OGSADQP];

- **semantic services** for discovering services and workflows, and managing metadata, such as: third party service registries and federated personalised views over those registries [LOR2003], ontologies and ontology management [WRO2003];

- services for supporting the **e-Science scientific method** and best practice found at the bench but often neglected at the workstation, specifically: provenance management [STEb] and change notification [MOR2003].

The final layer (e) constitutes the applications and application services that use some or all of the services described above.

9.1.2.2 Services that form the experiments

myGrid middleware must go hand in hand with corresponding development of domain specific scientific services that can deliver data and computation analysis.

Therefore bioinformaticians within the project have been developing service based access to bioinformatics tools and data.

- **Bioinformatics services**: Services such as database retrieval and analysis tools need to be wrapped and offered in a form that accommodates their distribution and variety of data formats. myGrid has acquired or wrapped a range of bioinformatics Web Services including: the complete EMBOSS application suite of over eighty

analysis tools, MEDLINE, SRS, OMIM and NCBI & WU BLAST sequence alignment tools. The project has developed

Soaplab[1], a universal connector for legacy command line based systems. The majority of services that we want to be able to make use of are shell scripts, PERL fragments or compiled architecture specific binaries rather than web services; Soaplab provides a fairly universal glue to bind these into web services and ia freely available.

- **Text extraction services:** AMBIT is a system for Acquiring Medical and Biological Information from text developed under the auspices of this and the CLEF e-Science project. The majority of biomedical knowledge still persists as free text in the published literature. More automated assistance in the delivery of this knowledge to the scientist requires at least some of the information is extracted into a more structured machine interpretable form. AMBIT provides an information extraction service based on natural language (http://industryhttp://industry.ebi.ac.uk/soaplab/)

9.1.2.3 Services for forming experiments

myGrid regards *in silico* experiments as distributed queries and workflows. Data and parameters are taken as input to an analysis or database service; then output is taken from these, perhaps after interaction with theuser, as input to further tools or database queries.

- **Workflow enactment, creation and management**:

  Once discovered or built, a workflow is run by our powerful FreeFluo[2] workflow enactment engine, which can handle WSDL based web service invocation.

  FreeFluo (http://freefluo.sourceforge.net) supports two XML workflow languages, one based on IBM's Web Service Flow Language (which we used early on) and our own, XScufl, developed as part of the Taverna project, in collaboration with the Human Genome Mapping Project [TAV2003]. The FreeFluo engine and the Taverna workflow development environment are open source and downloadable. See [ADDOIN] for further details.

- **Distributed database query processing**: The OGSA-DAI project (http://www.ogsadai.org/) and myGrid are together building a distributed query processing system that will enable a user to specify queries across a set of Grid-enabled information repositories in a high level language (initially OQL). Complex queries on large data repositories may result in potentially high response times, but the system can address this through parallelisation. The initial prototype is to be released in August 2003. See [OGSADQP] for further details.

- The **myGrid Information Repository** (mIR) acts as a personalised store of all information relevant to a scientist performing an *in silico* experiment. It implements an information model tailored to e-Science.

  Experimental data is stored together with provenance records of its origin. It is used to store workflow specifications ready to be submitted to the enactor together with records of running or completed workflows. These workflow records form a major basis for internally generated data provenance. The mIR has also been designed to store information about people and projects both [3]directly linked to the investigation and from the wider scientific community to aid collaboration.

Metadata storage is a central feature of the mIR, with annotation possible for all internally stored objects in addition to objects stored in disparate remote repositories. Annotations are currently stored in an RDF triple like manner (http://www.w3c.org/RDF/) and the project is considering the use of ”off the shelf” RDF triple stores such as the Jena Semantic Web toolkit (http://www.hpl.hp.com/semweb/jena.htm). Several types of annotation are used from free-text notes of the object’s significance with respect to the investigation, to more structured DAML+OIL based ontology annotations of what the object represents [HOR2002]. Annotation is a key tool used to link related objects and so answer wideranging queries such as ’What workflows have been recently run by members of my project?’ and ’What other data is available on this topic?’

9.1.2.4 Services for discovery and metadata management

Much of e-Science depends on discovering and pooling resources especially services but also experimental designs, data, people and projects. myGrid has developed several components to facilitate this discovery process.

- **Registries and registry views.** These are a key feature of web services infrastructure in which service descriptions are centrally published. myGrid extends the idea of a registry in three ways [GOB2002]:
  - *Personalised views over distributed registries*. It has become clear that multiple distributed registries will exist, some community wide, some specific to an organisation. To accommodate this registry views are been developed that aggregate distributed information based on a personal profile.
  - *Extensible metadata storage*. Originally designed to support the web services standard UDDI, the registry has now been underpinned with a flexible RDF storage component which enables it to support additional metadata standards such as DAML-S and BioMOBY.
  - *Additional semantic descriptions* to allow more precise searching by both people and machines. These DAML+OIL semantic descriptions build on the work of the DAML-S coalition (http://www.daml.org/services) and have been used to guide the construction of workflows by constraining the choice to those services, which have semantically compatible inputs and outputs. Similarly semantic description of workflows has been used within the myGrid workbench to discover relevant workflows given an item of data selected from the mIR.

9.1.2.5 Services for supporting e-Science

myGrid aids users in finding appropriate resources, offering alternatives to busy resources and guiding users in the composition of resources into workflows. In addition, myGrid offers:

- **Notification services**: A workflow may need to be re-run when new or updated data and analytical software become available. myGrid has a notification service to mediate an asynchronous interaction between services. Servers may register the type of notification events they produce and clients may register their interest in receiving updates. Notifications may also be used to automatically trigger workflows to analyse new data. See [MOR2003] for further details.
- **Provenance management**: When a workflow is executed, FreeFluo generates provenance logs in the form of XML files, recording the start time, end time and service instances operated in this workflow. Data, and metadata about the workflow and the

provenance logs are stored in the mIR. All mIR objects carry provenance attributes, hence the provenance log has who created it, when, in what context, and so on.

In addition, a set of metadata is associated with this workflow invocation instance: the input and output relationships between the workflow instance and data items, the 'is defined by' relationship between the workflow instance and its associated definition documents. Other annotations regarding the hypothesis of the experiment, thoughts and opinions by the scientist and quality of results are also stored as XML in the mIR or as regular web documents. This provenance information is extracted to answer questions such as "what recent workflows were run by Dr. Pearce using BLAST". As myGrid makes liberal use of ontologies, by annotating provenance logs with concepts drawn from the myGrid ontology, it is possible to expand experiment with building a dynamically generated hypertext of provenance documents, data, services and workflows based on their associated concepts and reasoning over the ontology [ZHA2003]. See [STEb] for further remarks on provenance.

- **Personalisation opportunities**: The intention of the project was to make all services personalised to the scientist in their own appropriate ways. For example: different users can be provided with appropriate views of the mIR; the registry view gives a user perspective over the services they can use, and the opportunity to attribute their metadata to third party services they do not own, as well as publish their own workflows and services; and the event notification system allows users to define their own choice of events. The user is represented by a User Proxy service.

9.1.2.6 Discussion

The current exploration of Graves' disease has been quite narrow and orientated around a single user group. To fully evaluate the myGrid components, population with data and associated metadata from multiple studies, investigations, projects, experiments and users is required.

Other components of myGrid need more significant ongoing development following lessons learnt from early prototypes. The initial investigations have revealed the need for a more sophisticated model of provenance and other experimental data holdings [WIL2002]. This will allow the storage of much more heavily linked metadata about provenance that will enable the creation of views of the mIR along many axes.

The myGrid Information Repository has provided the project with many useful insights into the types of data and metadata that need to be stored and the ways that data needs to be presented. The current mIR uses RDBMS technology and much of the information held therein is stored in a triple like manner. Much of the provenance information is stored as XML files; this makes it cumbersome to retrieve and process much of the metadata stored in the mIR. Consequently, the project will be investigating more wide spread use of RDF technology in the future.

Currently, the notification service is coarse grained in the types of notifications it indicates. For instance, one topic is "data change", which is used for the arrival of new data, the update of data, etc. in the mIR.

9.1.3 BRIDGES: Biomedical Research Informatics Delivered by Grid Enabled Services

The information with respect to the BRIDGES project is taken fromits web site. The project and its results are briefly presented here because it experimented and provided intial

implementation of (a) access services to public databases, (b) the project worked on the development of a Grid service which provides a parallelised BLAST service for the users and (c) worked on the utility of various existing and emerging federation tools and the use of OGSA-DAI. All these areas are very important and relevant to the ACGT architecture and workplan. Lessons learned by others are therefore a very welcomed starting point.

BRIDGES is developing and exploring database integration over six geographically distributed research sites within the framework of the large Wellcome Trust biomedical research project Cardiovascular Functional Genomics. Three classes of integration were developed to support a sophisticated bioinformatics infrastructure supporting: data sources (both public and project generated), bioinformatics analysis and visualisation tools, and research activities combining shared and private data. The inclusion of patient records and animal experiment data means that privacy and access control are particular concerns. Both OGSA-DAI and IBM Information Integrator technology have been employed.

The project was aiming at delivering the following results:

- An effective environment for biomedical bioinformatics supporting the work of the Wellcome Trust Cardiovascular Functional Genomics project. This will include federated access to data, analysis and visualisation across at least the UK centres with appropriate authorisation and privacy.
- An improved understanding of the requirements for the support of academic biomedical research virtual organisations. This was published as a final project report and exemplified with publicly available re-usable data access and integration components.
- An evaluation of the utility of various existing and emerging federation tools (e.g. replication tools such as GIGGLE, query tools such as DiscoveryLink and platforms such as OGSA-DAI) in this class of application.

The results of the BRIDGES project are of particular relevance to ACGT, because similar areas are addressed in the ACGT Workplan. Specifically, the BRIDGES project evaluated the OGSA-DAI specifications as compaterd with IBM Information Integrator technologies. Their experiences are very useful for the work in ACGT related to the selection of technologies for Data Access. Also, the project experimented with the implementation of Grid based analytical toolls. Although the tools required in ACGT are quite different, the experiences aquired in BRIDGES (with the GT3) in this domain are very relevant for ACGT.

9.1.3.1 Data Integration - Public Domain Data

To allow the integration of data from heterogeneous public data sources, BRIDGES experimented with the use of two different technologies (for the purpose of their evaluation/comparison). IBM's Information Integrator is a commercial package that allows the binding of data sources through standardised wrappers specific to the type of data source. The OGSA-DAI package achieves the same end by using Grid services as a layer between client applications and the data source. Users can browse the data via the GeneVista visualisation tool, which is available via the BRIDGES portal.

One surprising outcome of the BRIDGES project has been the lack of programmatic access to live biological databases. This has introduced an additional requirement for a data warehouse which is populated with data derived from flat file data dumps of the public domain databases. Data federation operates across the warehouse and the few available databases with programmatic access.

9.1.3.2 Computational Resources – GridBLAST

Like many other biomedical research projects, the CFG project has a need for large scale biological sequence comparisons. These are mostly carried out using the BLAST tool, a widely used search algorithm that is used to compute alignments of nucleic acid or protein sequences with the goal of finding the n closest matches in a target data set. This is computationally costly and therefore a task that can benefit from being adapted to run on a compute Grid.

The project has developed a GT3 based Grid service which provides a parallelised BLAST service for the users. Multiple query sequences are partitioned into sub-jobs on the basis of the number of idle compute nodes available and then processed on these in batches. To achieve this, we have written our own java based scheduler which distributes sub-jobs across an array of resources.

For enabling the control of which users access which resources (which may be necessary depending on the local policies at individual resources) the project has implemented a role based access control system which uses the PERMIS Grid authorisation software (http://www.permis.org/). For each job submitted, our service queries a PERMIS authorisation service for the roles a given user has, and allocates resources according to these.

Job submission to the National Grid Service is through GSI-enabled Globus 2 jobs, with Java Cog kit client side code, and uses a host proxy for authentication. This eliminates the need for our users to acquire and manage digital certificates and instead user authentication happens at the BRIDGES portal by means of standard username and password pairs. The PERMIS software gives us the option of then controlling what users are allowed to do once authenticated.

The client side code that was used for job submission to the NGS is publicly available on its web site (http://www.brc.dcs.gla.ac.uk/projects/bridges/public/code.htm).

9.2 References

[TAV2003] Taverna workflow environment for bioinformatics (http://sourceforge.net/projects/taverna), Proceedings UK OST e-Science 2nd All Hands Meeting, September 2003.

[ADDOIN] M Addis, T Oinn, M Greenwood, J Ferris, D Marvin, P Li, and A Wipat. Experiences with e-Science workflow specification and enactment in bioinformatics. In [2].

[OGSADQP] MN Alpdemir, A Mukherjee, NW Paton, P Watson, AAA Fernandes, A Gounaris, and J Smith. OGSA-DQP: A Service-Based Distributed Query Processor for the Grid. In [2].

[GAIHEP] R Gaizauskas, M Hepple, N Davis, Y Guo, H Harkema, A Roberts, and I Roberts. AMBIT: Acquiring Medical and Biological Information from Text. In [2].

[GOB2003] CA Goble, S Pettifer, and R Stevens. Knowledge Integration: In silico Experiments in Bioinformatics in The Grid: Blueprint for a New Computing. Morgan Kaufman, 2003.

[GOB2002] CA Goble and D De Roure. An Application of the Semantic Web. ACM SIGMOD Record, 31(4), December 2002.

[HOR2002] I Horrocks. DAML+OIL: a Reason-able Web Ontology Language. In Proc. of EDBT 2002, number 2287 in Lecture Notes in Computer Science, pages 2–13. Springer, March 2002.

[LOR2003] P Lord, C Wroe, R Stevens, CA Goble, S Miles, L Moreau, K Decker, T Payne, and J Papay. Semantic and Personalised Service Discovery. In S Miles, J Papay, V Dialani, M Luck, K Decker, T Payne, and Luc Moreau. Personalised Grid Service Discovery. Performance Engineering. In Performance Engineering. 19th Annual UK Performance Engineering Workshop, pages 131– 140, 2003.

[MOR2003] L Moreau, X Liu, S Miles, A Krishna, V Tan, and R Lawley. myGridNotification Service. In T Oinn. Talisman Rapid Application Development for the Grid. In Proceedings of Intelligent Systems in Molecular Biology, Brisbane, Australia, July 2003.

[PARWAT] S Parastatidis and PWatson. The NEReSC Core Grid Middleware. In M Senger, P Rice, and T Oinn. Soaplab - a unified Sesame door to analysis tools. In [2].

[STEa] R Stevens, K Glover, C Greenhalgh, C Jennings, P Li, M Radenkovic, and A Wipat. Performing in silico Experiments on the Grid: A Users Perspective.

[STEb] R Stevens, M Greenwood, and CA Goble. Provenance of e-Science Experiments – experience from Bioinformatics.

[STE2003] R Stevens, A Robinson, and CA Goble. myGrid: Personalised Bioinformatics on the Information Grid. Bioinformatics, 19:i302–i304, 2003.

[TAL2002] D Talia. The Open Grid Services Architecture - Where the Grid Meets the Web. IEEE Internet Computing, 6(6):67–71, 2002.

[WIL2002] MD Wilkinson and M Links. BioMOBY: an open source biological web services proposal. Briefing in Bioinformatics, 3:331–341, 2002.

[WRO2003] C Wroe, R Stevens, C Goble, A Roberts, and M Greenwood. A suite of DAML+OIL ontologies to describe bioinformatics web services and data. International Journal of Cooperative Information Systems, 12(2):197–224, 2003.

[ZHA2003] J Zhao, CA Goble, M Greenwood, C Wroe, and R Stevens. Annotating, linking and browsing provenance logs for e-Science. In 2nd Intl Semantic Web Conference (ISWC2003) Workshop on Retrieval of Scientific Data, Florida, USA, October 2003.

10 Management Systems and Standards for Clinical Trials

10.1 Introduction

Over the last years, getting patients and doctors into clinical trials and the lack of open-source software for the management of the clinical data and combining them with data from research labs are the most important aspects that delay the process of getting better treatments for patients with cancer. Whereas great care is needed in the development of a clinical trial recruitment program, a lot of effort has to be put into the development of Clinical Trial management systems.

In this regard an Electronic Data Capture (EDC) in the management of clinical data is a computerized system designed for the collection of clinical data in electronic format for use mainly in human clinical trials.

Typically, EDC systems provide:

- a graphical user interface component for data entry
- a validation component to check user data
- a reporting tool for analysis of the collected data

EDC systems are used by Life Sciences companies, broadly defined as the pharmaceutical, medical device and biotechnology industries in all aspects of clinical research, but are particularly beneficial for late-phase (phase III-IV) studies and pharmacovigilance and post-market safety surveillance. EDC can increase the data accuracy and decrease the time to collect data for studies of drugs and medical devices.

EDC had its origins in another class of software—Remote Data Entry, or RDE—that surfaced in the life sciences market in the late 1980s and early 1990s. Clinical research data—patient data collected during the investigation of a new drug or medical device—is collected by physicians, nurses, and research study coordinators in medical settings (offices, hospitals, universities) throughout the world. Historically, this information was collected on paper forms which were then sent to the research sponsor (e.g., a pharmaceutical company) for data entry into a database and subsequent statistical analysis environment. This process has a number of shortcomings, though:

- data is copied multiple times, which produces errors
- errors that are generated are not caught until weeks later
- visibility into the medical status of patients by sponsors is delayed

To address these and other concerns, RDE systems were invented so that physicians, nurses, and study coordinators could enter the data directly at the medical setting. By moving data entry out of the sponsor site and into the clinic or other facility, a number of benefits could be derived:

- data checks could be implemented during data entry, preventing some errors altogether and immediately prompting for resolution of other errors

- data could be transmitted nightly to sponsors, thereby improving the sponsor's ability to monitor the progress and status of the research study and its patients

These early RDE systems used "thick-client" software—software installed locally on a laptop computer's hardware—to collect the patient data. The system could then use a modem connection over an analog phone line to periodically transmit the data back to the sponsor, and to collect questions from the sponsor that the medical staff would need to answer. With the rise of the Internet in the mid 1990s, RDE became a web-based software representing a new class of EDC software.

The development of RDE systems was pushed by the establishing of standards for the management of clinical trials. In the 1980s the European Union began harmonising regulatory requirement. In 1989, Europe, Japan, and the United States began creating plans for a common harmonisation. The **International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use** (ICH) was then created in April 1990 at a meeting in Brussels. ICH is a project that brings together the regulatory authorities of Europe, Japan and the United States and experts from the pharmaceutical industry in the three regions to discuss scientific and technical aspects of pharmaceutical product registration.

The purpose of ICH is to reduce or obviate the need to duplicate the testing carried out during the research and development of new medicines by recommending ways to achieve greater harmonisation in the interpretation and application of technical guidelines and requirements for product registration.

Harmonisation would lead to a more economical use of human, animal and material resources, and the elimination of unnecessary delay in the global development and availability of new medicines while maintaining safeguards on quality, safety, and efficacy, and regulatory obligations to protect public health.

Six parties that represent the regulatory bodies and the research-based industries of the founding members are responsible for the decision making process: the European Union, the European Federation of Pharmaceutical Industries and Associations, the Ministry of Health, Labour and Welfare, the Japan Pharmaceutical Manufacturers Association, the Food and Drug Administration, and the Pharmaceutical Research and Manufacturers of America.

Further standards in data management are described in chapter 11.6.

10.2 Data flow in Clinical Trials (SIOP 2001 / GPOH)

Looking from the point of a clinician into a clinical trial, it is obviously that the patient with his disease is in the center of consideration. In contrast to Figure 1, in this section the ACGT environment will be outlined from the perspective of the patient. All of the data are only existing, because there is a patient having a disease or a tumour and he is enrolled in a clinical trial, receiving treatment or delivering his tumour to a research unit. All research and as a consequence all data generated in the lab or in the clinic are based on this fundamental issue. Figure 9 reveals this.

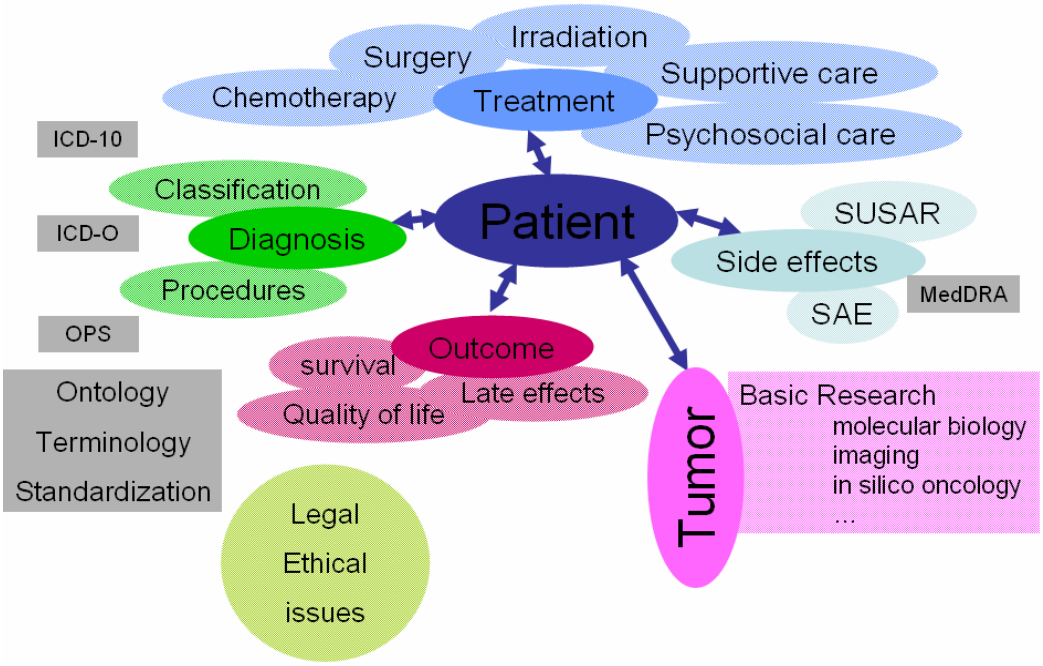


Figure 12: Outiline of a clinical trial from the perspective of the patient

For the Nephroblastoma trial (SIOP 2001 / GPOH) the data flow was worked out and is shown in Figure 10.

The interpretation of the data flow for the trial allows to define scenarios in clinicogenomic trials that are necessary on a granular level and raises a number of questions, that have to be addressed.

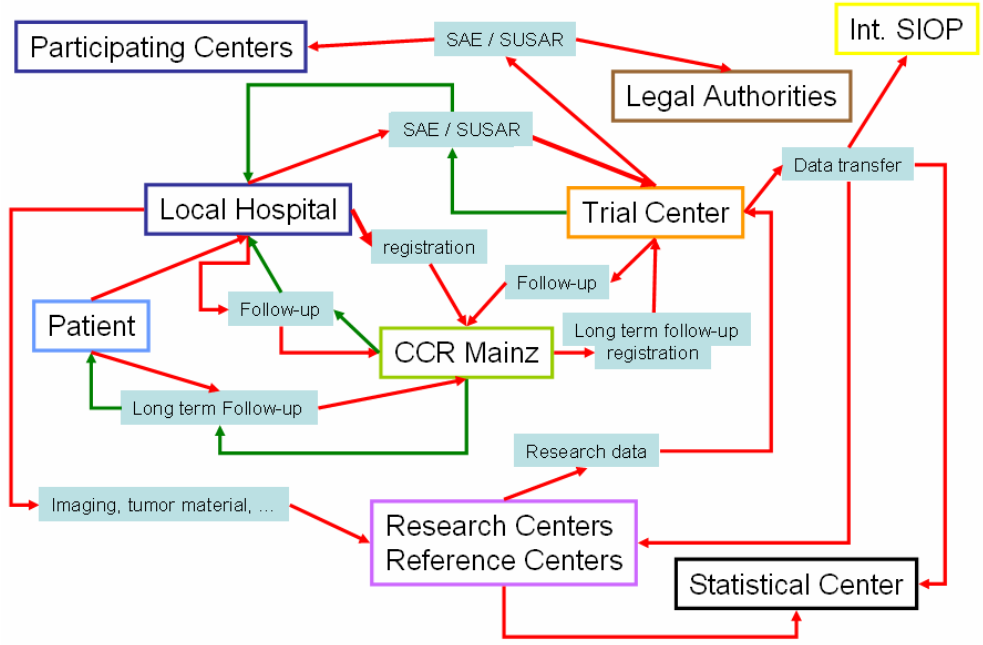


Figure 13: Data flow of the SIOP 2001/GPOH trial)Red arrows illustrate data that are send, green arrwos illustrate request for data; CCR: Childhood Cancer Registry

The following list shows scenarios that can be defined:

- Anonymisation of personal data
- Pseudonymisation of personal data
- Electronical signature of data
- Reporting of SAEs and SUSARs
- Transferring of DICOM data (imaging studies, etc.)
- Export and import of data (registration, trial center, statistical center, etc.)
- …

The following questions have to be addressed:

- What data have to be anonymised ?
- What data have to be pseudonymised ?
- What data have to be signed ?
- Must there be a feedback after sending of relevant data ? (for example for the treatment of the patient)
- Who has access to which data ?
- How to handle data protection and data security ?
- How to handle signatures ?
- …

The list of scenarios and questions has to be completed by People of workpackage 10, 11 and 12.

10.3 Data flow in Clinico-Genomic Trials

The data flow of the SIOP 2001/ GPOH trial can be easily transferred to other clincal trials. The integration of the data flow between research centers and knowledge based Web services builds a model for the data flow in ACGT. Figure 11 shows a general model for the data flow in ACGT.

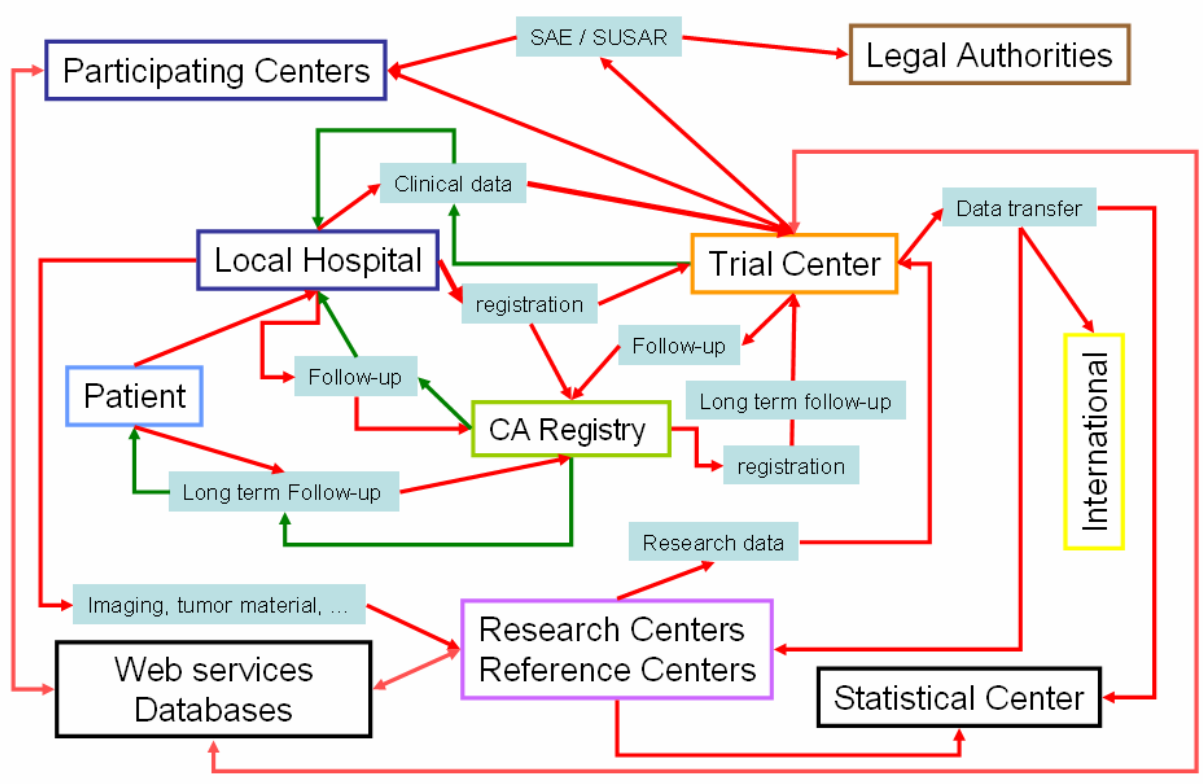


Figure 14: General data flow model in ACGT. (Red arrows illustrate data that are send, green arrows illustrate request for data; CA Registry: Cancer Registry)

The data flow in clinical trials has to be independent of time, meaning that data can be send and received everytime.

10.4 Data Management Systems in Clinical Trials

"Clinical Data Management (CDM) is the process of collecting and validating clinical information with the goal of converting it into an electronic format to answer research questions and to preserve it for future scientific investigation." [CHO2004] CDM is a crucial part of the clinical trial process, which ensures the validity, quality, integrity and completeness of the collected data. Researchers often underestimate the difficulty and importance of data management. But statistics can be only as good as the data they are based on.

According to [CHO2004] the CDM process includes:

- Case report form development
- Database development
- Data entry, query and correction
- Data quality assurance
- Data lock; archive and transfer.

Today more and more software systems are emerging that are able to support most aspects of Clinical Data Management. The requirements for those Clinical Data Management Systems are to support all steps of the CDM process, provide comfortable usability for all participating parties and to be compliant to the regulations /guidelines. We have evaluated

currently available Clinical Data Management Systems and present the results in this document

First the most important aspects of the Clinical Data Management process will be described. The following chapters are dedicated to relevant regulatory requirements, guidelines and standards. The last section describes features of currently available Clinical Data Management Systems and the evaluation of these systems.

10.5 Clinical Data Management

10.5.1 Case report form development

A case report form [CRF] is a printed or electronic form that is designed to collect the required research and administrative data from a subject in a clinical trial. The measurement and recording of the trial data are a very critical step and have direct impact on the quality of the data collected for a clinical trial. It is therefore important that the CRFs are designed with clarity and ease of use in mind.

Ideally the following principles should be considered when designing CRFs:

CRFs should be designed to capture all and only the data required per the study protocol.

CRFs do not only determine **what data** has to be collected but also **in what order** it has to be collected and **who** collects it.

They should be designed to collect data elements in standardized format to ease data entry. An appropriate coding of the data is also necessary to facilitate further data analysis. Standard modules for CRFs should be used wherever possible. They are developed to collect the same data across different clinical trials and using them leads to the development of standard procedures, such as data entry, saving time and effort. Standard modules are often stored for their reuse in standard libraries. Redundant data elements on the CRFs should be avoided, e.g. collecting the age as well as the birth date. A set of well developed CRFs is not completed without a guide that describe how to fill in the CRFs. [CHO2004, RON2000]

10.5.2 Database development

After designing the data entry forms the clinical trial database has to be developed considering the structure of the CRFs and the requirements specified in the study protocol. The trial database has the purpose to store all data that is relevant for the trial, i.e. the data captured with the CRFs and other trial specific data.

The trial database should provide easy access to data for later statistical analysis as well as for administrative purposes during the conduction of the trial (e.g. report functions like determining the number of patients per clinic during trial conduction should be implemented).

The data captured in the database should be standardized according to the CRFs. The database should comprise comprehensive metadata to assure that the stored data can be later used in a sensible way. Standardizing the data among different clinical trials (using standardized database modules) is desirable since that makes the results of those trials

comparable. During implementation data security aspects have to be considered (e.g. audit trails, data backup).

Designing the case report forms and trial database are crucial for the success of a clinical trial. Errors made in these phases can result in very cost intensive consequences in later stages of the clinical trial. [CHO2004, RON2000]

10.5.3 Data entry and correction

In many multicentric clinical trials today still paper based CRFs are used. The CRFs are filled in from the participating hospitals and then they are sent to a central data facility. There the data is monitored and entered from data entry clerks into the clinical trial database. This is normally accomplished by double data entry (entry by two different data entry clerks and validating the entries against each other) to minimize transmission errors. For any missing or inconsistent information a Query Form has to be filled in for clarification. This is a very time consuming process often resulting in sending data back and forth.

Therefore, today Remote-Data-Entry-Systems (RDE-systems) are emerging where the data are captured at the participating site and transferred electronically to the trial central data facility. RDE has the advantage that the data can be validated during data capture (e.g. logic checks can be performed) and the system can give alerts on incorrect or missing data. This results in increased data quality and double data entry and the procedure that the data entry clerks have to query missing or inconsistent data is unnecessary.

Since it is not easy to shift directly from paper based CRFs captured at a central data facility to RDE there exist also “solutions in between”. These systems rely on “fax-based data recognition” and “automated data acquisition from optical images” techniques. These approaches can be relatively easy integrated into current working practice since the data is also captured on paper based CRFs. The CRFs are then faxed or scanned and can be read by automatic recognition techniques directly into the trial database.

Data Correction

As required by ICH GCP any changes or corrections to a CRF have to be dated, initialed, and explained (if necessary) and should not hide the original entry. [CHO2004, RON2000]

10.5.4 Data quality assurance

Data validation is the cleaning of trial data in order to assure that they attained a reasonable quality level. The ICH GCP requires that the sponsor is responsible that quality control should be applied to each stage of data handling during a clinical trial to ensure that all data are reliable and have been processed correctly.

There are different data quality inspection procedures for tracking errors of data in the database. Some examples are edit checks, source data verification or double data entry.

The requirement for data monitoring of a Clinical Trial is defined in the ICH GCP (see section 3.1.) as the act of overseeing the progress of a clinical trial, and of ensuring that it is conducted, recorded and reported in accordance with the protocol, standard operating procedures (SOP) Good Clinical Practice, and applicable regulatory requirement(s). The trial monitor performs source data verification. It is required by ICH GCP and means that the monitor of a trial has to check the CRF entries with the source documents (original medical records) and assure that all data are correctly and completely recorded and reported.

To ensure completeness and consistency of data, data validation or plausibility checks are usually programmed and run on the clinical trial database:

Edit check specifications are developed according to CRFs and the study protocol. The following should be considered when reviewing the consistency check specifications: missing or incomplete data, range checks included to detect potentially invalid data, within-subject consistency-checks. Queries should be generated whenever inconsistencies are detected and send to the trial monitor for clarification. [CHO2004, RON2000]

10.5.5 Data lock, archive and transfer

When the database is complete, i.e. there are no outstanding queries, signatures of the responsible individuals are required to finalize (or lock) the database. After the database lock, a written approval from the responsible facility is required to initiate any changes to a finalized database. The final locked database should remain active in the system for at least three months before it is archived.

After the database is closed the statistical analysis is performed. For this purpose the relevant data is extracted from the database into formats that can be processed from statistic analysis programs (e.g. SAS, SPSS).

Database transfer at a specific format can be done per request, e.g. from the sponsor [2, 9].

10.6 Standards in Data Management

10.6.1 Clinical Data interchange standards consortium (CDISC)

CDISC is an open, non-profit organization that has the mission to develop and support platform-independent data standards that enable information system interoperability to improve medical research. CDISC seeks to achieve its mission through the development of standard data models designed to support the end-to-end data flow of clinical trials. These support the electronic acquisition, exchange, submission and archiving of clinical trials data and metadata.

CDISC has developed a model that specifies a format for interchange and archive of data, the Operational Data Model (ODM) as well as several content models that define standards for datasets for Data submission (SDS), standards for laboratory data (LAB) and standards for analysis datasets (ADaM). An overview of the CDISC models is given in figure 12.

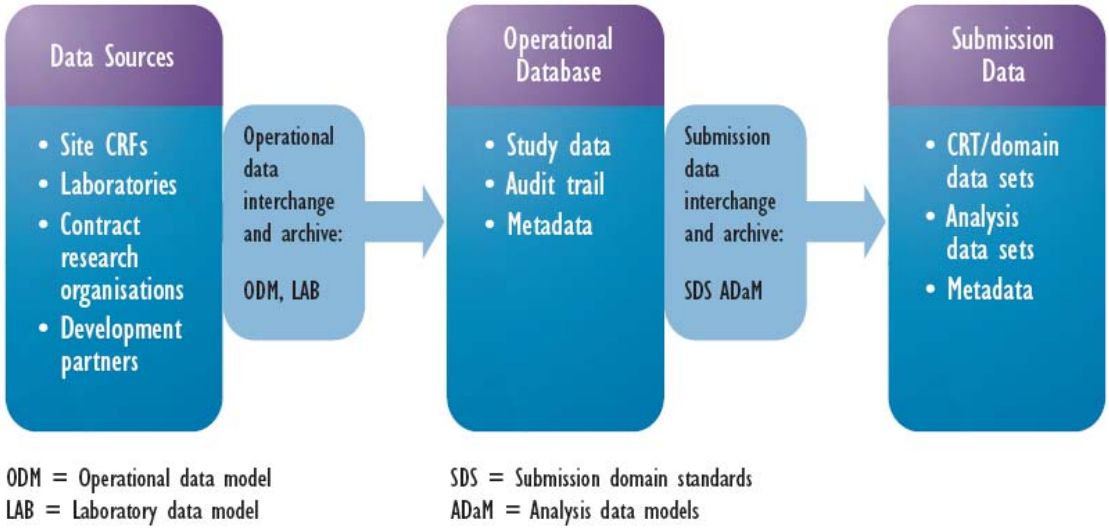


Figure 15: The SCDISC models (taken from[KUS2003])

The models are briefly described in the following:

Operational Data Model (ODM)

The Operational Data Model (ODM) is a platform independent format for interchange and archive of data collected in clinical trials. The model represents study metadata, clinical data and administrative data associated with a clinical trial. Information that needs to be shared among different software systems during a trial, or archived after a trial, is included in the model. In March 2006 version 1.3 was published.

Laboratory Data (LAB)

The LAB model is a standard model for the interchange of clinical trial laboratory data. It is a content model, which can be implemented through different means, including ASCII, SAS, XML and HL7 Version 3 RIM messaging.

Submission Data Standards (SDS)

The Submissions Data Standards are a set of standards developed by CDISC that are intended to guide the organization, structure, and format of standard clinical trial tabulation datasets submitted to a regulatory authority. Since 2004 the Study Data Tabulation Model (SDTM) is the standard format of the FDA that sponsors of human drug clinical trials can use to submit data to the agency.

Analysis Dataset Model (ADaM)

The Analysis Dataset Model (ADaM) is a set of guidelines and examples for analysis datasets used to generate the statistical results for submission to a regulatory authority such as FDA. It specifically addresses needs of statistical reviewers.

Protocol Representation

The Protocol Representation working group aims to define a model to present the trial protocol. They have identified standard elements (including a subset of those in SDTM) of a

clinical trial protocol. It is an HL7 initiative with a CDISC team to support the standards development activities. A machine-readable model is an ultimate goal for this team.

10.6.2 Terminology working group of CDISC

CDISC is also working in cooperation with representatives of government, academia and pharmaceutical companies to define standard codelists for use in the clinical research data interchange. This work will support other CDISC standards. The terminology that has been developed until now has been loaded into NCI Thesaurus (National Cancer Institute), one of NCI's premier vocabulary products. It is intended to integrate the terminology into the NCI Cancer Data Standards Repository or caDSR, an ISO 11179 metadata repository.

Today sponsors have been implementing individual models (using ODM to export and import data or building their databases on SDTM) but currently no guidelines exist that describe how to effectively use all of the collective models in combination. So the aims for the future are to harmonize and further improve the models. CDISC wants to achieve final harmonisation of its models until 2010. CDISC cooperates with HL7 and has many sponsors from the pharmaceutical industry. The FDA also supports the standards of CDISC. Therefore the importance of CDISC is permanently increasing. [CDI2006]

10.6.3 Terminologies in Clinical Trials

Terminologies are used in Clinical Trials to standardize the coding of data. Appropriate and unambiguous coding is essential for a good data quality. There are a lot of medical terminologies in existence e.g. the ICD (International Classification of Diseases) to encode diseases or LOINC (Logical Observation Identifiers Names and Codes) for encoding laboratory observations that are used today in clinical trials. Here we want to focus on MedDRA, since this terminology has been adopted by the ICH as the standard medical terminology for regulatory communication.

MedDRA - the Medical Dictionary for Regulatory Activities

MedDRA is a medical terminology that was designed to be applicable to all phases of drug development, excluding animal toxicology. Before the introduction of MedDRA a variety of often not compatible terminologies developed by different organizations were used at different stages of clinical trials. That often led to misunderstandings and loss of information. The need to standardize the medical terminology used internationally for regulatory affairs of medical products was recognized by regulatory agencies as well as pharmaceutical companies. A Working Group of representatives from the ICH and an observer for the World Health Organization worked to develop the MedDRA terminology and completed version 1.0 in 1994. The ICH Steering Committee adopted it as an international medical terminology for regulatory purposes.

MedDRA is based on the UK Medicines and Healthcare products Regulatory Agency (MHRA) medical terminology and it incorporates different other medical terminologies (e.g. WHO-ART, ICD, ICD-CM, COSTART and J-ART).

It includes terminology for symptoms, signs, diseases and diagnoses. In addition, it contains the names of investigations (e.g. Liver function analyses, metabolism tests), sites (e.g. application site reactions, implant site reactions and injection site reactions), therapeutic indications, surgical and medical procedures and medical, social and family history terms. It does not include a drug or device nomenclature or terms covering study design, pharmacokinetics or patient demographics.

The terminology does not contain definitions for its terms. Synonymous terms are grouped under one standard term called preferred term. These preferred terms are ordered under a five-level hierarchical framework. FDA and the European Agency for the Evaluation of Medical Products (EMEA) recommend the use of MedDRA in their regulations.

It is maintained and distributed by the MedDRA Maintenance and Support Services Organization (MSS0). It is available at a reasonable cost, and it will be updated at a frequency that is appropriate to the needs of users. [CHO2004, MED2006, RON2000]

10.7 Clinical Data Management Systems

Powerful data collection and management systems for clinical trials exist. We have analyzed how well the systems support the described clinical data management procedures, the support of standards, the compliance to regulations, the usability and the price of the systems.

In the following an overall overview about the features of the analyzed Clinical Data Management Systems shall be given, in Appendix A description of the single systems can be found.

1. **Case report form development**
   All but one of the analyzed Data Management Systems provide tools to design and layout CRFs. That can be done from clinicians without informatics skills but in every case one or two-day training courses are necessary.

2. **Database development**
   The database is normally generated automatically from the CRFs. When designing the CRFs the user is almost free in defining the items for the CRFs and the names for the database tables and attributes. That is problematic since that can result in ambiguous names that lack standardization.
   All systems are client server architectures.

3. **Reporting functions to support administrative purposes during trials**
   Most of the systems provide a good support to generate different kind of study management reports and patient data listings

4. **Data entry**
   All of the described systems support RDE. One of the system supports additionally fax based data entry (DataFax).

5. **Data quality assurance - Support for monitoring activities**
   Some of the described Data Management Systems have features to support the tasks of the clinical monitors. In most systems it is possible to add queries of the monitors to every item of the CRF. These questions are shown to the investigator and can be answered by them. Some systems include messaging systems that allows the participating parties to interact with each other.

6. **Data quality assurance – Support for data edit checks**
   Specifying complex validation checks and conditions for individual questions on the eCRFs is possible with all of the systems.

7. **Data export**
   In all systems data can be exported in different formats (e.g. in SAS or SPSS or CDISC ODM format)

8. **Usability**
   The companies of all the systems claim that their systems are very user friendly. Since we could not test the systems in most case we are not able to decide if it is true what the companies write.

9. **Price**
   Prices of the commercial systems are mostly dependent on different aspects (e.g. the number of users, the used modules, duration of the trial, etc) only one of the analyzed systems is provided under an open source license (openClinica).

10. **Compliant to regulations for clinical trials**
    All of the reviewed systems are compliant to ICH GCP and FDA 21 CFR part 11.

11. **Use of standards**

Since all of the systems are compliant to ICH GCP and FDA 21 CFR part 11 all of them provide advanced security mechanisms (password protected access, audit trail, encrypted data transmission,..).

All have an advanced roles and rights management (Add new users, assigning user permissions and defining roles).

For clinical trials that are sponsored by research organizations often in-house developments are used not at last because of the fact that commercial clinical data management systems are very expensive. We have not included those systems into our analysis. [WEB2004]

In the following section the systems analyzed are summarized.

10.7.1 Review of Data Management Systems for Clinical Trials

Overview of Clinical Data Management Systems

In the following table a list of currently available Clinical Data Management Systems is shown. We have sent a questionnaire to all of the companies querying the features of the systems. We have integrated only the systems of the companies that answered our request into our analysis.

Name	Company	URL
itrial	Interactive Clinical Technologies Incorporated	http://www.icti-almac.com
TrialXS	Clinsource	http://www.clinsource.com
eResearch Network	eResearch Technology	http://www.ert.com
eClinical Suite	Etrials Worldwide, Inc.	http://www.etrials.com
Datafax	Clinical DataFax Systems Inc.	http://www.datafax.com

secuTrial	iAS	http://www.secutrial.com
MACRO	InferMed	http://www.infermed.com
Oracle Clinical	Oracle	www.oracle.com
InForm	PhaseForward	http://www.phaseforward.com
StudyWorks	PHT Corporation	http://www.phtcorp.com
TRIALink	TeamWorks	http://www.teamworks.de
Studymanager	ACS	http://www.clinicalsoftware.net
openClinica	Akaza Research	http://www.openclinica.org

Clinical Data Management Systems - Features

White spaces in the tables indicate that the information is not known to us.

secuTrial

Name	**secuTrial**
Company	IAS
General	
URL	www.secutrial.com
Advantages according to the company	flexible, user friendly, web based, high experience of the company; modular and scalable structure, combining all components in one user interface, intelligent and intuitive user guidance;
Technical support necessary?	Depending on configuration
Price	Depending on number of patients, duration of trial
Features	
Data entry	RDE
CRF development	FormBuilder can be used to create Case Report Forms
Edit checks	To define arbitrarily detailed consistency checks and interdependencies is possible; Form logic and edit checks:

	able to checks the form's completion grade on submit of data; Format checks: Checks logical conditions (numbers, dates etc.) Plausibility check: Univariat (compare to limits for the specific field) Multivariate (compare to related values in other fields)
Support of data monitoring	For every data item in a CRF a monitor can enter a question that can be answered by the investigator; Message system that allows participants of the study to communicate with each other
Reporting	Concise reports possible, different report and analysis can be set up (for example showing graphical view of how many patients have been recruited)
Data export	Export of data is possible in the following formats: SAS, CDISC-XML, CSV or TXT
Supported standards	CDISC XML
Regulatory compliance	ICH GCP, 21 CFR part 11
Technology	Database: Oracle 9i

MACRO

Name	**MACRO**
Company	InferMed
General	
URL	www.infermed.com
Advantages according to the company	Audit trail, electronic signatures, clinical query management
Technical support necessary?	Depending upon situation; Obligatory annual maintenance fee that covers:

	- Software functional error correction - Incident logging - Minor upgrades (approximately every 6 months) - Significant discount on major upgrades (min 30%)
Price	Depending on number of users and their role
Features	
Data entry	RDE; Multimedia Data collection is fully supported
CRF design	CRFs can be designed by no technical user; but a 1-2 days training is necessary,
Edit checks	Validation rules for the eCRFs can be set up in the “expression editor”
Support of data monitoring	Communication with data monitors via a system of electronic notes attached to individual questions on the CRFs
Reporting	Creating of self defined data views possible Querying of databases possible
Data export	Extraction of data in the following formats is possible: SAS, SPSS or CSV
Supported standards	MedDRA; CDISC ODM import and export
Regulatory Compliance	ICH GCP, 21 CFR part 11, EC Clinical Trial Directive
Technology	Database: MSDE 2000, SQL Server 2000, Oracle 8.1.7

etrials

Name	**eClinical Suite**
Company	Etrials Worldwide, Inc.
General	
URL	http://www.etrials.com
Advantages according to the company	- 450 clinical trials, top 20 pharmaceutical/ biotechnology companies, SOPs, QA Audits, input of different data - Only fully integrated platform for the collection, management and

	analysis of clinical data
Technical support necessary?	Training and support are most important
Price	Depending on number of users and their role
Features	
Data entry	RDE
CRF development	Training is necessary but can the be performed from a user without informatics skills
Edit checks	Built-in edit checks
Reporting	End users can access a robust ad hoc reporting tool from within the application
Monitoring	Monitors are able to apply queries or notes to any data item on a CRF
Data export	Export possible in Excel spreadsheets, XML, various delimited text files,
Supported standards	Supports CDISC SDTM and ODM formats; Etrials supports integration with standard industry dictionaries (COSTART, MedDRA, ICD) and customer-specific dictionaries
Regulatory Compliance	ICH GCP, 21 CFR part 11, Health Insurance Portability and Accountability Act (HIPAA)
Technology	Database: Oracle

DataFax

Name	**DataFax**
Company	Clinical DataFax Systems Inc.
General	
URL	http://www.datafax.com
Advantages according to the company	Compliant with FDA, to have paper CRF for using FAX
Technical support necessary?	License fee includes technical support during office hours

Price	1 user license: $ 16.000/year, 2 users: $ 29.000, 2nd year 50 %
Features	
Data entry	Paper based CRFs send by fax and can be recognized automatically; In the next version RDE will be also possible
CRF development	CRF design possible
Edit checks	Edit checks can be integrated (provides powerful edit check language)
Support of data monitoring	
Reporting	Includes standard study management reports, patient data listings, workflow summaries, data queries, descriptive statistics, etc.
Data export	Export in ASCII files, SAS datasets and Oracle, PostgreSQL and MySQL possible
Supported standards	
Regulatory Compliance	ICH GCP and 21CFR part 11
Technology	Database: own internal database structure (indexed ASCII files)

StudyManager

Company	ACS
General	
URL	www.clinicalsoftware.net
Advantages according to the company	Most widely used, most mature, best supported
Technical support necessary?	Does not require assistance from informatics when the trial is running
Price	Depending on modules, anticipated users, implementation
Features	
Data entry	RDE
CRF	Easy-to-use design and layout tool let create CRFs to capture virtually

development	any type of data,
Edit checks	A wide variety of validation rules can be assigned to each data field, ensuring that data is entered correctly. Ranges of acceptable values can be set or whether data is required within a field. Multiple rules for a single field can be set and field values can be validated against other fields
Support of data monitoring	
Reporting	Flexible report builder that enables customization of report templates
Data export	
Supported standards	Accepts all types of data coding, based on the costumers requirements
Regulatory Compliance	ICH GCP and 21 CFR part 11, HIPAA Compliance
Technology	Database: Microsoft SQL; Frontend is web based application; uses .net technology

InForm

Name	**InForm**
Company	PhaseForward
General	
URL	http://www.phaseforward.com
Advantages according to the company	> 1000 clinical studies across 100 countries
Technical support necessary?	Not necessary
Price	Depending on number of CRF and patients, duration of trial
Features	
Data entry	RDE
CRF development	eCRFs have to be done by the company at the beginning
Edit check	

Support of data monitoring	
Reporting	
Interfaces for analysis tools/ data export	
Supported standards	
Regulatory Compliance	ICH GCP; CFR Part 11
Technology	Database: Oracle

openClinica

Name	**openClinica**
Company	Akaza Research
General	
URL	www.openclinica.org
Advantages according to the company	Leading open source clinical trials management system
Technical support necessary?	No, but the company offers users to subscribe to support services that provide software users with the assurance of professional technical assistance in a timely and thorough manner (not for free)
Price	Open Source LGPL licence
Features	
Data entry	RDE
CRF development	CRFs can be designed by using Excel Design Templates
Edit checks	For every data item on the CRF an validation expression (regular expressions, range checks..) can be defined and an error message that should appear if an entered value does not satisfy the validation criteria
Support of data monitoring	No special support

Reporting	Limited reporting functions, possible to define datasets and filters
Interfaces for analysis tools/ data export	Datasets can be created including the data collected on different CRFs, this datasets can be exported in currently 4 different formats: HTML, tab-delimited text, comma-delimited text, or SPSS format
Supported standards	CDISC
Regulatory Compliance	ICH GCP; CFR Part 11
Technology	

10.8 References

[CDI2006] Clinical Data Interchange Standards Consortium. http://www.cdisc.org, last accessed: 24.04.06

[CHO2004] Chow S, Liu P (2004). Design and Analysis of Clinical Trials. Second Edition, JohnWiley & Sons, Hoboken, New Jersey.

[EUD2001] DIRECTIVE 2001/20/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL (2001). http://eudract.emea.eu.int/docs/Dir2001-20_en.pdf. last accessed: 24.04.2006

[COM2005] COMMISSION DIRECTIVE 2005/28/EC (2005). http://pharmacos.eudra.org/F2/eudralex/vol-1/DIR_2005_28/DIR_2005_28_EN.pdf. last accessed: 24.04.2006

[FDA2006] FDA Food and Drug Administration. http://www.fda.gov/. last accessed: 24.04.2006

[ICH2006] ICH International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. www.ich.org. last accessed: 24.04.2006

[KUS2003] Kush R (2003). The world of Standards for Clinical Research, http://www.touchbriefings.com/pdf/16/Kush.pdf, last accessed: 24.04.2006

[MED2006] MedDRA and the MSSO. http://www.meddramsso.com/MSSOWeb/index.htm, last accessed: 24.04.2006

[RON2000] Rondel R, Varley S, Webb C (2000). Clinical Data Management. 2. Aufl, John Wiley & Sons, Chichester.

[WEB2004] Weber, Ralf (2004). Terminologiebasierte Erstellung von rechnerunterstützten Dokumentationssystemen in klinischen Studien; Dissertation Medizinische Fakultät Heidelberg.

[GAL2005] Galea-Lauri, Joanna; Forster, Louise (2005). An overall Guide to Interventional Clinical Trials of Medicinal Products for Researchers at ICH/GOSH [GOSH ICH/05/S03/02 (Revised 18th March 2005)]. Approved by: Professor David Goldblatt (Clinical Director, Research and Development) http://www.ich.ucl.ac.uk/ich/r&d/ctguidelines.pdf, last accessed: 21.04.2006

[PRI2006] Pritchard-Jones, Kathy (2006). A brief questionaire about the implementation of the EU CTD and its impact on newly opening paediatric trials Kathy, UK, personal communication.

11 Tools for the creation and management of clinical trials

11.1 Introduction

Clinical Data Management (CDM) is the process of collecting and validating data during a clinical trial. Therefore it is a crucial part of the clinical trial process, which ensures the validity, quality, integrity and completeness of the collected data. It is still one of the main problems in that process that clinical data management systems often do not integrate extensive metadata to describe the data collected during a clinical trial. That makes reuse of the data difficult even for human experts. [CHO2004] For the services in the ACGT environment it is crucial that reliable, well defined clinical data is provided as their input.

A very important part of the clinical trial protocol is the definition of data that has to be recorded for every patient in the trial. Standardizing these data and unambiguous definitions is a prerequisite for consistent quality permitting reuse and sharing of data. Based on the clinical trial protocol, the documentation system is developed. The most important parts of such a system are the “Case Report Forms” (CRFs) which are designed to collect the required research and administrative data and the clinical trial database for storing these data.

In many multicentric clinical trials today still paper based CRFs are used. From the participating hospitals thousands of CRFs are sent to a central data facility. There the data is monitored and entered into the clinical trial database. This is very time consuming often resulting in sending data back and forth during this process. Therefore, today the preferred systems are Remote-Data-Entry-Systems (RDE-systems) where the data are captured at the participating site by a computer-based application system and transferred electronically to the trial central data facility.

Today there are a lot of powerful clinical data management systems in existence. Most of these systems allow creating eCRFs (electronic CRFs) without any informatics skills. According to the CRFs the clinical trial database is normally generated automatically. The user is free in defining the items for the CRFs and the names for the database tables and attributes. That is problematic since it often results in ambiguous names that lack standardization. Often terminologies can be used to code the data in these systems but not to define the metadata of trial database and CRFs. To analyse the clinical trial data that has been collected with these clinical management systems the data has to be exported and external statistical packages have to be used. The clinical data management systems normally allow data export in various formats (e.g. SPSS, SAS or CDISC-ODM). The relevant data for export has to be filtered manually.

ACGT has the vision of a distributed Grid environment to support cancer research. Data from different clinical trials can be shared and brought together for advanced analysis and data mining. Ideally analysis and knowledge discovery tools are able to access and understand the collected data automatically. Syntactic and semantic interoperability is necessary. This interoperability can only be achieved when the data is collected in a standardized form. The clinical trial database needs to have comprehensive machine and human understandable metadata that defines the syntactic as well as the semantic level of the data.

To achieve semantic interoperability ontologies will play a major role in ACGT, e.g. the ACGT ontologies will be used to integrate legacy databases into the Grid environment.

Although it is not the aim of ACGT to implement a clinical data management system the project nevertheless also wants to explore how development and functionality of such systems can be enhanced by basing them onto formal ontologies especially focusing on interoperability. A vision for the future is e.g. that a person without informatics skills could create standardized eCRFs that allow storing the data in clinical trial databases with comprehensive metadata without the help of software developers. This problem can be solved through basing the trial protocol and the documentation system on a formal ontology. Formal ontologies comprise logical descriptions that serve to computationally define terms as well as human-readable definitions. Therefore, they are understandable by humans and by machines. The use of clinico-genomic ontologies that will be shared in the Grid environment will define the semantics of the data that is important for an accurate understanding and further processing of the data.

In the following we will review relevant work. The first chapter will be dedicated to research projects that also aim to create standardized clinical trial management systems by exploiting the use of ontologies or terminologies. The second chapter will describe CDISC, an organization that provides standard data models for clinical trials. Semantic Web and Ontology driven architecture that enhances model driven architecture focus also on the use of ontologies as metadata. Therefore, the last chapter will give a short summary of these topics.

11.2 Ontology based creation of documentation systems for clinical trials

11.2.1 caBIG and ISO/IEC 11179 metadata repository

The aims of caBIG (cancer Biomedical Informatics Grid) are similar to those of ACGT. caBIG is a network of individuals and institutions that are working to create a better environment for the sharing of cancer research data and software tools (www.cabig.org). The National Cancer Institute (NCI) developed the Cancer Common Ontologic Representation Environment (caCORE) to provide a framework for creating syntactically and semantically interoperable biomedical information services to achieve interoperability across the systems it develops as well as the systems developed by the caBIG program.

caCORE aims to achieve semantic interoperability through harmonization of information models, metadata and ontology/terminology. Therefore caCORE provides three layers of semantics by integrating the caBIO module (Cancer Bioinformatics Infrastructure Objects), the caDSR (Cancer Data Standards Repository) and the EVS (Enterprise Vocabulary Server). In the following we will describe the EVS and caDSR in more detail.

EVS is a description-logic based thesaurus and ontology management system. It provides ontology development and hosting services. Beside other terminologies and ontologies it contains the NCI Thesaurus and the NCI Metathesaurus.

The NCI provides a metadata repository called caDSR that is a comprehensive set of metadata descriptors for cancer research. This metadata repository uses the ISO/IEC 11179 model for metadata registration. The purposes of this standard are to provide a common

understanding of data across organizational elements and between organizations and the reuse and standardization of data.

According to that standard a Common Data Element (CDE) is the basic container for data. Each CDE is composed of two parts, a Data Element Concept and a Representation. The Data Element Concept describes the semantic part of the CDE, it comprises an object type and a property (e.g. address and street). Both can be taken from a terminology or ontology. Therefore the caDSR provides an interface to the EVS for easy selection from all contained ontologies and terminologies. The Representation describes the form of the data containing the data type, a value domain and if necessary a unit of measurement. For a detailed description of the caDSR see [NCI].

The caBIG project contains a tool called FormBuilder that can be used to create CRFs. A form is simply a collection of CDEs assuring that all data collected with the form has comprehensive metadata. It is possible to assign questions to the CDEs that will appear on the form. Several CDEs can be grouped to form a template, template forms are generic forms that can be used as the basis for creating the actual forms. The created CRFs can be downloaded in different formats. The described mechanism assures that the data collected on this CRF are standardized.

The Cancer Standards Repository is very complex. It comprises more than 13.000 CDEs arranged into different contexts that refer to the facilities that created them. [CAD, PHI2006]

One missing piece of this infrastructure is the representation of object relationships in caDSR metadata. It is also not able to support formal ontologies and formal statements. That limits the possibility for automatic reasoning and inference on this data.

That issue is addressed in the XMDR (Extended Metadata Registry Project) initiative. Historically, metadata registries have been used to communicate semantic information about information artifacts among various human database and applications designers and developers (and sometimes end users). Developments in agents, inference engines, and web services technologies have generated increased interest in using ISO/IEC metadata registries to convey semantics information for use by other programs (machine processing). Such uses require more precise and more formal descriptions of semantics.

XMDR is concerned with the development of improved standards and technology for storing and retrieving the semantics of data elements, terminologies, and concept structures in metadata registries. The NCI participates in this project among other organizations. A prototype XMDR is under development [XMDR]. For some of their suggestions see [OLK2005a, OLK2005b]. So in the future it is possible that the EVS and the caDSR are merged to an XMDR.

11.2.2 openEHR archetypes

OpenEHR is a non-profit foundation developing open specifications and open-source software for interoperable, life-long electronic health records. The initiative was started in 1992 as an EU research project entitled “Good European Health Record” and was later continued under the name “Good Electronic Health Record”.

The most remarkable concept developed by openEHR is the two-level methodology that has been introduced to create future-proof health information systems by separating the level of knowledge from the level of information.

In traditional Health Information Systems domain concepts are hard-coded into the software and database models. Since medicine is a domain with a large number of concepts and a high rate of concept changes these systems do not have a long lifetime and are very expensive to extend and maintain.

In the two-level approach software and database schemas are developed based on a reference information model. This model only comprises the non-volatile concepts that do not change over time. Domain experts define the level of knowledge during runtime of the health information system by creating archetypes. Archetypes present the volatile domain concepts and can be defined by constraining the concepts of the information model.

An example for a concept of an information model is Observation. An example for an archetype that constrains the concept Observation is blood pressure. [BEA, TZE2004]

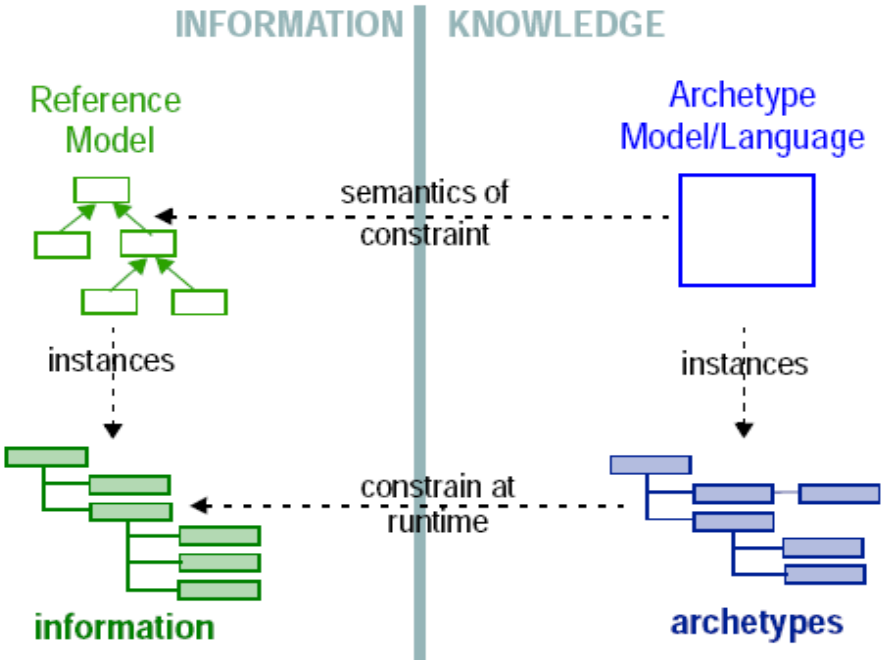


Figure 16: Overview of the openEHR two level modelling approach for EHRs (taken from [BEA])

OpenEHR defines a special language to construct archetypes, the “Archetype definition language” (ADL). An alpha version of an archetype editor is available that provides a GUI based comfortable development environment for clinicians [OCE].

When archetypes are used at runtime in particular contexts, they are composed into larger constraint structures, with local or specialist constraints added, via the use of templates. These templates can be directly translated into user interfaces. However, currently no user interface generator is provided publicly by the project.

Archetypes define the semantics of the data. For querying the archetypes openEHR is currently developing an Archetype Query Language (AQL) that will be based on paths, since archetypes are completely path-addressable in a manner similar to XML data, using path expressions that are directly convertible to Xpath expressions [OPA].

Two information systems are interoperable on the syntactic layer when they use the same reference information model. These information systems can exchange information in the

form of archetypes or templates. The archetype definitions ensure interoperability on the semantic level. In the future it is envisioned to share archetype definitions by storing them in public archetype repositories.

The archetypes enable domain experts to define structure, constraints and terms of the data and in a second step the user-defined terms can be bind to terms of an ontology or a terminology. It is currently not possible to create archetypes in the way that the archetype developer is authored of an underlying ontology (e.g. in OWL). But that would be of great advantage because this would guarantee that only sensible structures could be created. In 2002 T. Beale one of the main developers of openEHR stated in the technical mailing list of the project that that would be possible in the future but would take some time to evolve [OPE, OPA].

The openEHR archetype approach has never been used to create documentation systems for clinical trials but Garde et al [GAR2005] have recently stated that it is highly desirable to adopt it for this purpose.

11.2.3 TERMTrial

The software system TERMTrial was designed to support terminology based development of documentation systems for multicentric clinical trials. It consists of a component for the definition and management of terminology systems for cooperative groups of clinical trials and two components for the terminology-based automatic generation of trials databases and terminology-based interactive design of eCRFs.

The overall structure of the system is depicted in the following figure. The terminology management system (TMS) consists of two components the Common Terminology Module (CTM) and the Trial Specific Terminology module (TSTM). In the Common Terminology Module a reference terminology for a cooperative group of clinical trials can be defined and maintained. For each new clinical trial a trial specific terminology can be created based on the reference terminology. All partners have to agree on the reference terminology and it has to be guaranteed that it is used exclusively in the way it is represented in the TMS. An authorization is needed to change and maintain it. A trial specific terminology can be transformed with the Database Generating Component automatically into a schema for the clinical trial database. The Form Building Component allows interactive design of terminological consistent case report forms.

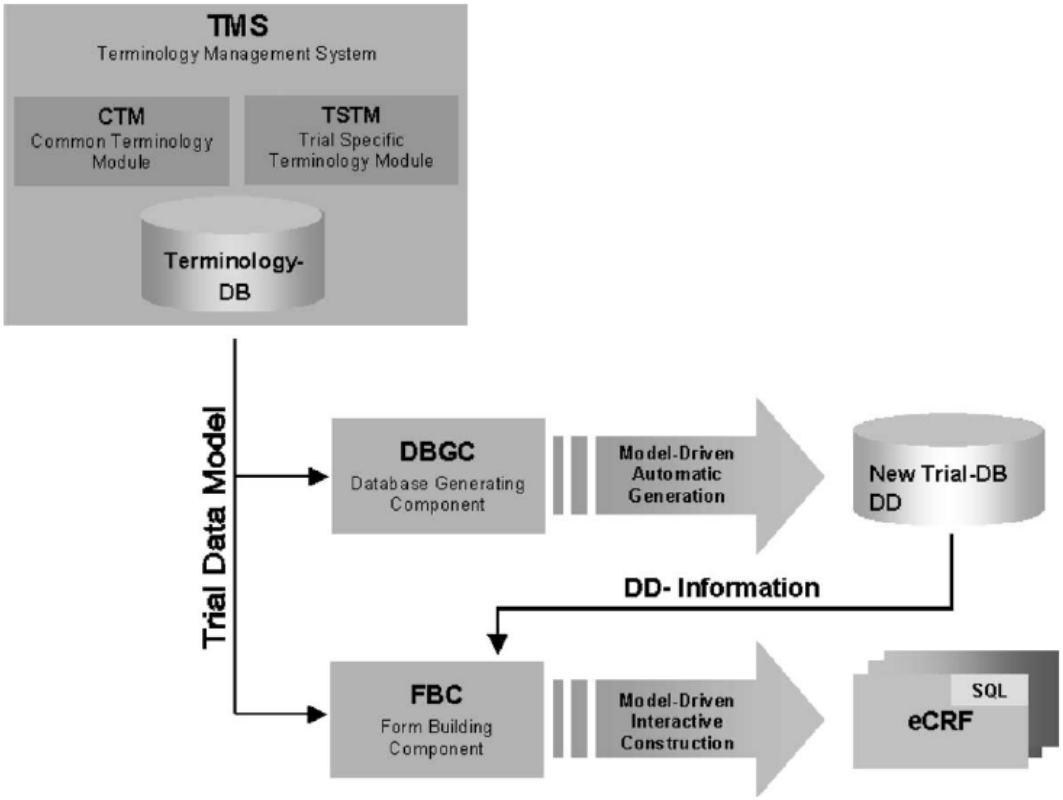


Figure 17: System Architecture of TERMTrial (taken from [MER])

In TERMTrial not only the clinical data but also the metadata of the databases is described by a centrally shared and maintained terminology since the database schema is generated from the terminology. That guarantees that the description of the data is standardised in all cooperating clinical trials and results and data are comparable. The generated databases are interoperable on the syntactic and semantic level.

The Database Generating Component and the Form Building Component have never been used in practice mainly due to the end of the project funding [MER2005, KNA2005].

11.2.4 Ontology based data dictionary for clinical trials

Heller et al. have developed a data dictionary for medical and trial-specific terms in which concepts and relations are defined. It can be provided to different medical research networks by means of the software tool “Onto-Builder” via the internet. The data dictionary is based on domain-specific ontologies and a top level ontology called “GOL”. The concepts and relations described in the data dictionary are represented in natural language, semi formally or formally according to their use [HEL2004].

In figure 15 an overview of the use of the data dictionary during the definition of a clinical trial is shown. However, this project has not provided tools for automatic database generation or CRF creation.

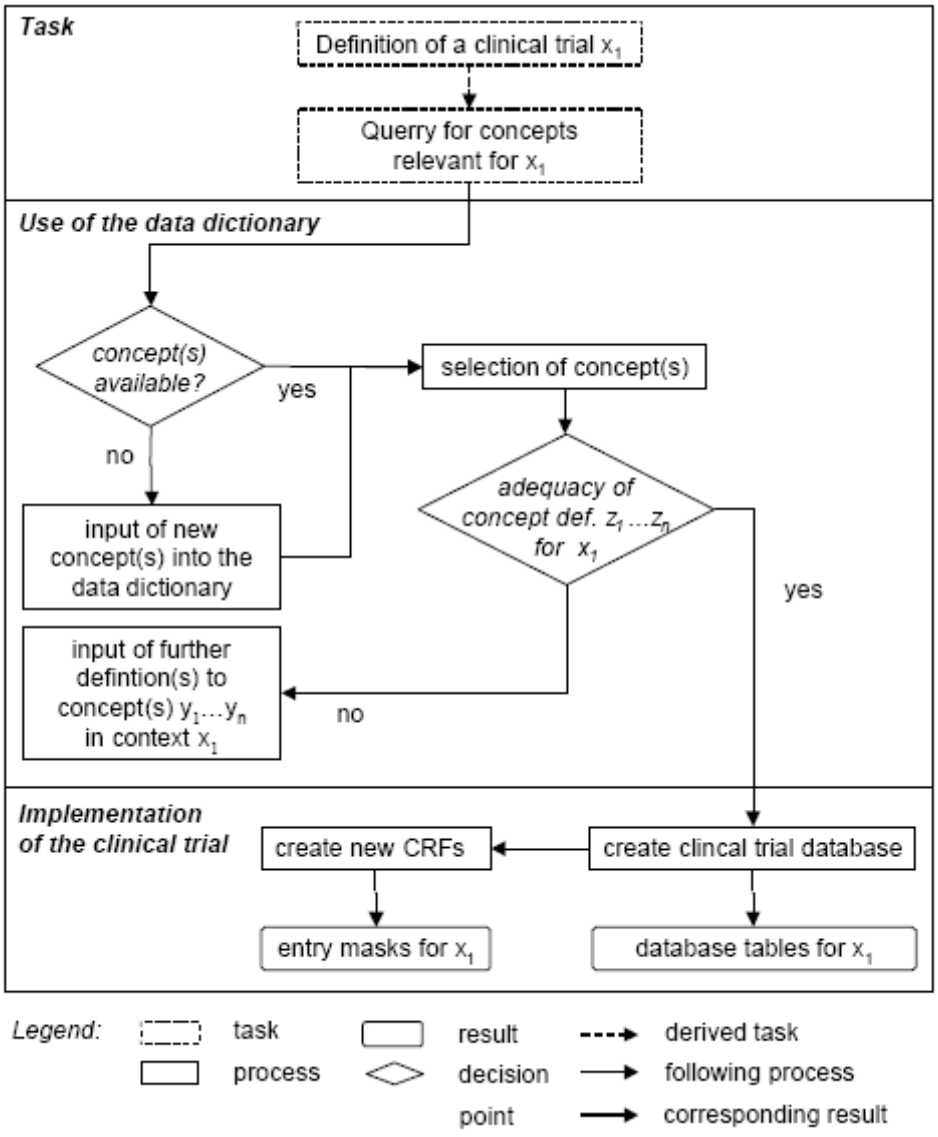


Figure 18: Use of the data dictionary in the clinical trial definition process (taken from [HEL])

The advantage of the two approaches described last is that a shared terminology or data dictionary minimizes inconsistencies occurring in the documentation and analysis of clinical trial data. Furthermore, the explicitly described basic concepts permit the unification of the meaning and interpretation of relevant medical concepts and clinical trial data.

However an disadvantage is that experiences in various medical fields have shown that the terminology-based approach is limited to specialized fields and it is argued that a comprehensive terminology is simply too complex and too difficult to maintain [GAR2005].

11.3 Ontology Driven Architecture and Semantic Web

The vision behind the Semantic Web is to make web-content machine-understandable so that it can be analysed by software agents and shared among Web Services. For that purpose the World Wide Web Consortium (W3C) is recommending a number of web-based languages that can be used to formalize web-content for describing machine understandable

metadata. Techniques of the Semantic Web are more and more integrated in practices for systems and software development.

While MDA provides a powerful and proven framework for Systems and Software Engineering, Semantic Web technologies can naturally extend it to enable representation of unambiguous domain vocabularies, model consistency checking, validation, and new capabilities that leverage increased expressivity in constraint representation. Ontologies will augment the OMG standards and methodologies giving rise to Ontology Driven Architecture (ODA).

The vision of Ontology Driven Architecture is that developers would discover sharable domain models from a variety of interrelated repositories and then build based on them their application. All applications that share overlapping domain models would then have a certain degree of interoperability built in. While this is still mostly a vision some promising approaches are beginning to appear [KNU2006, TET2006].

The Semantic Web community has produced a set of complimentary languages and tools for developing, maintaining, using and sharing domain models for Software Engineering, amongst other purposes. At the core are the languages OWL (Web Ontology Language) and RDF (Resource Description Framework) Schema, OWL being optimized to represent structural knowledge at a high level of abstraction. OWL is founded on Description Logics. This underlying formal logic makes it possible to exploit intelligent reasoning services such as automatic classification and consistency checking. These services can be used at build-time and therefore facilitate the construction of reusable, well-tested domain models. Reasoning services can also be used at runtime for various purposes. For example, this makes it possible to define classes dynamically, to re-classify instances at runtime and to perform complex logical queries. In addition to their foundation on logics, OWL and RDF Schema operate on similar structures to object-oriented languages, and therefore can be effectively integrated with traditional software components. Domain models in any of these languages can be uploaded and linked into the Web and application developers may be able to locate a suitable model on the web and simply reuse it [KNU2006, TET2006].

In this context the OBO (Open Biomedical Ontologies) foundry has to be mentioned. This project is a collaborative experiment, involving a group of ontology developers who have agreed to the adoption of a growing set of principles specifying best practices in ontology development.

The primary objective is to establish gold standard reference ontologies for individual domains of inquiry. Their goal is to develop a set of ontologies which can be used in combination because they are based on common principles. They state that the methodology of developing application ontologies always against the background of a formally robust reference ontology framework, and of ensuring updating of application ontologies in light of updating of the reference ontology basis, can ensure the interoperability of application ontologies constructed in its terms [OBO].

11.4 Discussion

The described projects show that ontologies are well suited for defining the underlying semantics of metadata. Basing domain models for software applications on reusable and shared reference ontologies can enhance the interoperability of these applications.

The approaches of caCORE to harmonize metadata, information model and ontology are of great importance for ACGT since they also aim to provide semantic interoperability in a Grid environment for cancer research. That there is still a lot to do and the approaches can be further enhanced is addressed in the XMDR initiative.

XMDR is cooperating with the OMG in an effort to exploit the semantic expressiveness of ontologies and integrating them into metadata and information models. They try to find ways how reference ontologies can be distributed and accessed in standard repositories to make them available for software developers and ease the integration in software. These approaches are heavily supported by the growing number of technologies emerging for the Semantic Web and try to make the vision of human and machine understandable metadata come true. But all of these techniques are still in their infancy and have not proved their usability in practice. Therefore exploring ontology driven architecture for clinical data management systems seems to be of high relevance for ACGT, to allow semantic interoperability of these systems and other services in the ACGT environment in the future.

It is also of great importance to cooperate with the CDISC organization. Since CDISC has strong support from industry its importance is permanently increasing. CDISC is not finalized yet and it is still under heavy development. Therefore ACGT can also give suggestions how their models can be improved and the expressiveness of ontologies can further enhance this standard.

11.5 References

[BEA] T. Beale. Archetypes: Constraint-based domain models for future-proof information systems. In: Eleventh OOPSLA Workshop on Behavioral Semantics: Serving the Customer (Seattle, Washington, USA, November 4, 2002). Edited by Kenneth
Baclawski and Haim Kilov. Northeastern University, Boston, pp. 16-32, 2002.

[CABIG] caBIG- cancer Biomedical Informatics Grid; https://cabig.nci.nih.gov (last accessed: 27.07.2006).

[CAD] caDSR – cancer Data Standard Repository; http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr (last accessed: 27.07.2006).

[CHO2004] Chow S, Liu P (2004). Design and Analysis of Clinical Trials. Second Edition, JohnWiley & Sons, Hoboken, New Jersey.

[EIC2005] M. Eichelberg, T. Aden, A. Dogac, G. B. Laleci. A Survey and Analysis of Electronic Healthcare Record Standards, ACM Computing Surveys, 37(4): 277-314, 2005.

[GAR2005] S. Garde, P. Knaup, T. Schuler, E. Hovenga. Can openEHR archetypes empower Multi-centre Clinical Research? In: Engelbrecht R, Geissbuhler A, Lovis C, Mihalas G (Eds.), Studies in Health Technology and Informatics 116, Connecting Medical Informatics and Bio-Informatics, Proceedings of MIE2005 (Geneva, 28.-31. August 2005): 971-976.

[HEL2004] B. Heller, H. Herre, K. Lippoldt, M. Loeffler. Standardized Terminology for Clinical Trial Protocols Based on Ontological Top-Level Categories, In: Kaiser, K., Miksch, S., Tu, S.W. (eds.) *Computer-based Support for Clinical Guidelines and Protocols. Proceedings of the Symposium on Computerized Guidelines and Protocols.* (CGP 2004), 13.-14. April 2004. Prag. p. 46-60. Studies in Health Technology and

Informatics, Vol.101. Amsterdam: IOS-Press.

[HL7] Health Level Seven; http://www.hl7.org (last accessed: 31.07.2006).

[HL7v3] HL7 Version 3 Guide; http://gim.upv.es/hl7/html/welcome/environment/index.htm (last accessed: 31.07.2006).

[KNA2005] P. Knaup, S. Garde, A. Merzweiler, N. Graf, F. Schilling, R. Weber, R. Haux. Towards shared patient records: An architecture for using routine data for nationwide research, International Journal of Medical Informatics, 75(3-4): 191-200, 2005.

[KNU2006] H. Knublauch, D. Oberle, P. Tetlow, E. Wallace. A Semantic Web Primer for Object-Oriented Software Developers, W3C Working Group Note (work in progress), 2006; http://www.w3.org/TR/sw-oosd-primer (last accessed: 06.05.2006).

[KUS] Kush R. The world of Standards for Clinical Research, 2003,http://www.touchbriefings.com/pdf/16/Kush.pdf (last accessed: 24.04.2006).

[MER2005] A. Merzweiler, R. Weber, S. Garde, R. Haux, P. Knaup-Gregori. TERMTrial-terminology-based documentation systems for cooperative clinical trials; Computer Methods and Programs in Biomedicine 78, 11 – 24, 2005.

[NCI] The NCICB User Applications Manual; http://ncicbsupport.nci.nih.gov/sw/content/NCICBAppManual.pdf (last accessed: 27.07.2006).

[OBO] The OBO Foundry; http://obofoundry.org (last accessed 06.05.2005)

[OCE] Ocean Informatics; http://oceaninformatics.biz/CMS/index.php (last accessed: 27.07.2007).

[ODM] Ontology Definition Metamodel; http://www.omg.org/ontology (last accessed 06.05.2005).

[OLK2005a] F. Olken, K. D. Keck, J. L. McCarthy. Improved Relationship Modeling in ISO/IEC 11179; Whitepaper 2005; http://metadata-standards.org/metadata-stds/Document-library/Documents-by-number/WG2-N0851-N0900/WG2-N0874-XMDR-Whitepaper-on-relationship-modeling-in-ISOIEC-11179.htm (last accessed 06.05.2005).

[OLK2005b] F. Olken, K. D. Keck, J. L. McCarthy. Ontologies and Formal Statements for ISO/IEC 11179; Whitepaper 2005; http://metadata-standards.org/metadata-stds/Document-library/Documents-by-number/WG2-N0851-N0900/WG2-N0873-XMDR-Whitepaper-on-Ontologies-and-Formal-Statements-for-ISOIEC-11179-discussions.htm (last accessed 06.05.2005).

[OPA] T. Beale, S. Heard, D. Kalra, D. Lloyd. OpenEHR Architecture Overview, Revision 1.0.1; The openEHR Foundation, 2006 http://svn.openehr.org/specification/TRUNK/publishing/index.html (last accessed: 27.07.2006).

[OPE] openEHR; www.openEHR.org (last accessed: 27.07.2006).

[PHI2006] J. Phillips, R. Chilukuri, G. Fragoso, D. Warzel and P. A. Covitz. The caCORE Software Development Kit: Streamlining construction of interoperable biomedical information services; BMC Medical Informatics and Decision Making 2006, 6:2.

[POL2004] J. T. Pollock, R. Hodgson (2004). Adaptive Information: Improving Business Through Semantic Interoperability, Grid Computing, and Enterprise Integration, First Edition,

JohnWiley & Sons, Hoboken, New Jersey.

[ROA] CDISC Roadmap Discussion Document; http://www.cdisc.org/standards/CDISCRoadmapJan2006.pdf (last accessed: 27.07.2006).

[TET2006] P. Tetlow, J. Z. Pan, D. Oberle, E. Wallace, M. Uschold, E. Kendall. Ontology Driven Architectures and Potential Uses of the Semantic Web in Systems and Software Engineering, Editors Draft (work in progress), 2006; http://www.w3.org/2001/sw/BestPractices/SE/ODA (last accessed: 27.07.2006).

[TZE2004] T. Z. Tzelepis, M. Tsiknakis, D.G. Katehakis, S. C. Orphanoudakis. Design and Implementation of "Two-level" Clinical Information Systems, Based on Archetypes; Proceedings of the 2nd International Conference on Information Communication Technologies in Health; July 8-10, 101 – 106, 2004.

[XMDR] Extended Metadata Registry (XMDR) Project; http://www.xmdr.org/ (last accessed: 06.05.2006).

12 Tools and techniques for the analysis of biomedical data

Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by the scientific community. This deluge of genomic information has, in turn, led to an absolute requirement for computerized *databases* to store, organize, and index the data and for specialized tools to view and *analyze* the data. The tremendous interest in *bioinformatics*, a new discipline at the intersection of molecular biology and computer science, is fuelled by the excitement surrounding the sequencing of the human genome and the promise of a new era in which genomic research dramatically improves the human condition. Advances in detection and treatment of disease and the production of genetically engineered foods are among the most often mentioned benefits.

Bioinformatics is a fertile new area for interdisciplinary research as well as a source for innovative information science and technology development. It has already served as an inspiration for many biological metaphors in computing, and conversely, information and computation paradigms have become ubiquitous in molecular biology. Researchers at the frontiers of biology and informatics are developing and can be expected to increasingly develop very novel symbiotic forms of science and technology.

A definition of Bioinformatics

[National Center for Biotechnology Information (NCBI);
http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html]

Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline…There are three important sub-disciplines within bioinformatics: the development of new algorithms and statistics with which to assess relationships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and the development and implementation of tools that enable efficient access and management of different types of information.

In the sequel we highlight the most crucial molecular biology modelling tasks accompanied by the bioinformatics methods, techniques, systems and tools needed, and actually utilised, in tackling these tasks.

12.1 Gene Discovery

The question of how many genes are encoded in the human genome is not yet answered. Recent estimates are converging in the 25,000 - 30,000 range; it could be many years before we have the final answer. Not only can the start and end positions of a predicted gene be wrong, but exons can be missed entirely or wrongly predicted to exist. Increased reliability in the gene predictions can be achieved by keeping only those predictions that show significant similarity to known cDNA sequences, obtained from transcribed DNA. However, evidence

that a stretch of DNA is transcribed does not definitively show the stretch to be a gene. We do not know how efficiently cells control transcription and it seems likely that non-gene DNA sequences are transcribed relatively frequently. Nor do we know how well the cell identifies transcripts that cannot be translated into protein. Proteins that cannot serve any useful function could be made, but rapidly removed.

12.1.1 Gene Discovery Approaches

In the last twenty years the discovery of genes and other functional sites in DNA sequences has evolved rapidly. The functional sites to be identified are splice sites, start and stop codons, branch points, promoters and terminators of transcription, transcription factor binding sites etc. [GEL1995]. Local sites such as these are called *signals* and methods for detecting them may be called *signal sensors*. Genomic DNA signals can be contrasted with extended and variable length regions such as exons and introns, which are recognized by different methods that may be called *content sensors*. Excellent recent surveys have been given by [GEL1995], [FIC1996a, FIC1996b], [GUI1997], [CLA1997], [MIL1998], [KRO1998], [BUR1997, BUR1998a, BUR1998b], [STO2000], [MAT2002] and [PER2001].

- *Dynamic programming approaches.* Nearly all integrated gene finders use dynamic programming to combine candidate exons and other scored regions and sites into a complete gene prediction with maximal total score. A brief and lucid tutorial on this topic can be found in [KRO1998] and a more detailed exposition in [DUR1998]. Dynamic programming approaches find the candidate gene structure with the best overall score. The key to success in these methods is developing the right score function.

- *Linguistic approaches.* They define a statistical model of genes that includes parameters describing codon dependencies in exons, characteristics of splice sites (e.g. the parameters of a weight matrix for splice sites), as well as ``linguistic'' information on what functional features are likely to follow other features. The linguistic rules for what functional features follow what other features are expressed by the parameters of a Markov process on the hidden variables. For this reason, these models are called hidden Markov models, or HMMs.

- *Comparative approaches* use explicit comparisons to other, previously known genes, or auxiliary information such as expressed sequence tag (EST) matches. Examples are Genie [KUL1997], geneid [XU1997], GeneParser3 [SNY1995], and recent versions of Grail [XU1997a]. The program AAT [HUA1997] and new versions of Grail also take into account EST information [XU1997b].

- *Promoter identification approaches.* Promoter regions are found upstream of the transcription start site of genes and are rich in signals corresponding to transcription factor binding sites. Promoter recognition methods have been developed for searching in a genome for a previously defined consensus and for extracting a consensus from a set of sequences [VAN1999]. Weight matrix approaches have also been used to identify functional motifs in promoters utilizing information about the sequential occurrence of specific nucleotides in a motif, expressed in the form of a weight matrix. The use of artificial neural networks (ANN) has also been explored in the annotation of promoters [REE2001]. Initialization of the ANN topology with established, reliable and universally accepted weighted matrices has been also tested with very good results [POT2001].

- *Comparative gene prediction.* A major development in the last few years, as more fully sequenced genomes have accumulated, has been the use of sequence conservation between species as a powerful indicator of gene structure in systems like TWINSCAN [KOR2001] or SGP2 [PAR2003].

It is also worth mentioning at this point a gene prediction software system named GeneID that has been developed by IMIM and that has been recognised as having a relatively good performance should be acknowledged, as well as the gff2ps software for the visualization of genomic sequences, which has been used in several genome publications including the human genome.
[http://genome.imim.es/main/software.html; http://www1.imim.es/software/geneid/index.html]

12.1.2 microRNA Gene Discovery

As opposed to protein-coding genes, neither signals as ribosome-binding sequences or splicing donor or acceptor nor compositional features originating from codons of the genetic code can be used. The features of the common transcriptional mechanisms like promoters, termination and other processing signals are too weak. Due to the different biogenesis of miRNAs in plants and vertebrates, the approaches for these organisms differ slightly. Using a comparative genomics approach, Jones-Rhodes and Bartel identify both computational and experimental 23 new miRNAs and their targets for *Arabidopsis thaliana* [Jones-Rhoades *et al.* 2004], extending the set of 69 known miRNAs.

[GRA2003] predict miRNAs in the *C. elegans* genome using sequence conservation and structural similarity to known miRNAs and generated 214 candidates additionally to the 53 known miRNAs and14 of the candidates were experimentally verified.

The pipeline developed by [LAI2003] for finding novel miRNA genes in *Drosophila melanogaster* (fruitfly) is called 'miRseeker'. They suggest 48 novel miRNAs additional to the 24 known miRNAs, of which they experimentally verified 24. An interesting computational approach to ncRNA gene-finding using comparative genome sequence analysis is implemented in the program QRNA by [RIV2001], and applied to the genome of *Saccharomyces cerevisiae* (yeast) [MCC2003] identifying 92 candidate ncRNA genes of which 13 were experimentally verified.

12.1.3 Available Systems and Tools

Gene prediction methods

Genie: http://www.cse.ucsc.edu/~dkulp/cgi-bin/genie

GenScan: http://ccr-081.mit.edu/GENSCAN.html

HMMgene: http://www.cbs.dtu.dk/services/HMMgene/

GeneMark-HMM: http://genemark.biology.gatech.edu/GeneMark/hmmchoice.html

Veil: http://www.cs.jhu.edu/labs/compbio/veil.html

GenScan: http://genes.mit.edu/GENSCAN.html

FirstEF: http://rulai.cshl.org/tools/FirstEF/

GeneSplicer: http://www.tigr.org/tdb/GeneSplicer/gene_spl.html

Geneid: http://genome.imim.es/main/software.html ;http://www1.imim.es/software/geneid/index.html

AAT: http://genome.cs.mtu.edu/aat.html

FGENEH: http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html

Genlang: http://cbil.humgen.upenn.edu/~sdong/genlang_home.html

GeneParser: http://beagle.colorado.edu/~eesnyder/GeneParser.html

Glimmer: http://www.cs.jhu.edu/labs/compbio/glimmer.html

Grail: http://compbio.ornl.gov/

MZEF: http://www.cshl.org/genefinder

Promoter recognition

TESS: http://www.cbil.upenn.edu/tess
Berkley: http://www.fruitfly.org/seq_tools/promoter.html

Datasets

Single genes: ftp://www-hgc.lbl.gov/pub/genesets/
Annotated contigs: http://igs-server.cnrs-mrs.fr/banbury/index.html
Sanger worm genefinding: http://www.sanger.ac.uk/Projects/C_elegans/
UCSC Genome browser for several organisms: http://genome.cse.ucsc.edu

12.1.4 References

[BUC1990] Bucher, P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* 212:563-578, 1990.

[BUR1997] Burge, C. *et al.* Prediction of complete gene structures in human genomic DNS. *J. Mol. Biol.* 268:78-94, 1997.

[BUR1998a] Burge, C.B. *et al.* Finding the genes in genomic DNA. Curr. Opin. Struc. Biol. 8: 346-354, 1998.

[BUR1998b] Burge, C.B. Modeling dependencies in pre-mRNA splicing signals. In Salzberg, S., Searls, D. and Kasifs, S., eds. Computational Methods in Molecular Biology, Elsevier Science, Amsterdam. pp. 127-163, 1998.

[BUR1996] Burset, M *et al.* Evaluation of gene structure prediction programs. *Genomics*, 34(3):353-367, 1996.

[CLA1997] Claverie. J.M. (1997). Computational methods for the identification of genes in vertebrate genomic sequences. *Human Molecular Genetics*, 6(10):1735-1744, 1997.

[DUR1998] Durbin, R. *et al. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[FIC1996a] Fickett, J. Finding genes by computer - the state of the art. *Trends in Genetics*, 12(8):316-320, 1996.

[FIC1996b] Fickett, J. The gene identification problem -- an overview for developers. *Computers and Chemistry*, 20(1):103-118, 1996.

[GEL1995] Gelfand, M.S. Prediction of function in DNA sequence analysis. *Jour. Comp. Biol.*, 2(1):87-115, 1995.

[GRA2003] Grad, Y. *et al.* Computational and Experimental Identification of C. elegans microRNAs. Molecular Cell, 11:1253-1263, 2003.

[GUI1997] Guigo, R. Computational gene identification: an open problem. *Computers and Chemistry*, 21(4):215-222, 1997.

[HOF1994] Hofacker, I. *et al.* Fast folding and comparison of RNA secondary structures. Monatshefte f. Chemie, 125:166-188, 1994.

[HUA1997] Huang, X. *et al.* A tool for analyzing and annotating genomic sequences. *Genomics*, 46:37-45, 1997.

[JON2004] Jones-Rhoades, M. *et al.* Computational Identification of Plant MicroRNAs and Their Targets, Including a Stress-Induced miRNA. Molecular Cell, 14:787–799, 2004.

[KOR2001] Korf, I. *et al.* Integrating genomic homology into gene structure prediction. Bioinformatics 17 Suppl 1: 140-148, 2001.

[KRO1998] Krogh, A. Gene finding: putting the parts together. In M. J. Bishop, editor, *Guide to Human Genome Computing*, chapter 11, pages 261-274. Academic Press, 2nd edition, 1998.

[KUL1997] Kulp, D. et al. Integrating database homology in a probabilistic gene structure model. In R. B. Altman, A. K. Dunker, L. Hunter, and T. E. Klein, editors, Proceedings of the Pacific Symposium on Biocomputing, pages 232-244. World Scientific, New York, 1997.

[LAI2003] Lai, E. et al. Computational identification of Drosophila microRNA genes. Genome Biology, 4:R42, 2003.

[MAT2002] Mathe, C. Current methods of gene prediction, their strengths and weaknesses. Nuc. Acid Res., 30:4103-4117, 2002.

[MCC2003] McCutcheon, J. et al. Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics. Nuc.Ac.Res., 31:4119-4128, 2003.

[MIL1998] Milanesi, L. et al. Prediction of human gene structure. In M. J. Bishop, editor, *Guide to Human Genome Computing*. Academic Press, 2nd edition, 1998.

[PAR2000] Parra, G. et al. Geneid in Drosophila. Genome Res. 10(4): 511-515, 2000.

[PAR2003] Parra,G. et al. Comparative Gene Prediction in Human and Mouse, Genome Research 13:108–117, 2003.

[PER2001] Pertea, M. et al. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res . 29(5):1185-90, 2001.

[POT2001] Potamias, G. et al. Knowledge-based TDNN architectures for features recognition DNA sequences. In Procs International Joint INNS/IEEE Conference on Neural Networks - IJCNN, Washington DC, USA, vol. 2, pp. 2327-2332, 2001.

[REE2001] Reese, M.G. Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. Comput Chem, 26(1): 51-6, 2001.

[RIV2001] Rivas, E. et al.. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics, 2:8, 2001.

[SNY1995] Snyder, E. et al. Indentification of protein coding regions in genomic DNA. JMB, 248:1-18, 1995.

[STA1984] Staden, R. Computer methods to locate signals in nucleic acid sequences. NAR, 12:505-519, 1984.

[STA1990] Staden, R. Finding protein coding regions in genomic sequences. Methods in Enzymology, 183:163-180, 1990.

[STO2000] Stormo, G. Gene-Finding Approaches for Eukaryotes. GenomeResearch, 10: 394-397, 2000.

[VAN1999] Vanet et al. Promoter sequences and algorithmical methods for identifying them. Res. In Microbiology, 150:779-799, 1999.

[XU1997a] Xu, Y. et al. Automated gene identification in large-scale genomic sequences. *Journal of Computational Biology*, 4(3):325-338, 1997.

[XU1997b] Xu, Y. et al. Inferring gene structures in genomic sequences using pattern recognition and expressed sequence tags. In *Proceedings, 5th International Conference on Intelligent Systems for Molecular Biology*, pages 344-353, 1997.

[ZUK1999] Zuker, M. et al. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: RNA Biochemistry and Biotechnology (Ed.: Barciszewski et al.), 11-43, 1999.

12.2 Structure Prediction

To overcome the excessively time-consuming work of improving rough drafts of x-ray crystallographic structures, researchers have developed sophisticated computational techniques utilizing predictive and comparative methods to fashion a new protein structure. Ab initio methods use the physiochemical properties of the amino acid sequence of a protein to literally calculate a 3D-structure. Many methods rely on different potentials to optimize the lowest energy model. As opposed to determining the structure of an entire protein, ab initio methods are typically used to predict and model protein folds (domains). Having the hypothetical structure for a part of the protein that interacts with a ligand, can potentially hasten drug exploration research.

Also, the way in which the RNA chain folds upon itself is crucial for the function of structural RNAs and can be predicted to some extent using specialized software for the determination of base pairing energies for RNA duplex. Two basic tasks are in order: RNA *secondary* and RNA *tertiary* structure prediction, reviewed in [MAJ2001]. For a review covering also DNA tertiary structure prediction, see [BEV2000]. If one restricts the problem to the prediction of unknotted secondary structures, elegant dynamic programming algorithms combined with nearest-neighbour free energy parameters combine to give rigorous solutions to the problems of computing minimum free energy structures, close to optimal folding, and partition functions that yield exact base pair probabilities. The open question in this field is to put forward a model including pseudo-knots that allows predictions in reasonable time. Using evolutionary profiles from multiple sequence-alignments, the per-nucleotide accuracy for known base-pairs has reached 80% [JUA1999]. Though secondary structure is energetically the strongest component for RNA structures, a theoretical limit is being approached in this field.

- *Dynamic programming minimization of free energy*. Dynamic programming minimization of free energy RNA secondary structure prediction based on free energy rules for stacking and loop formation is one of the major breakthroughs in the field of structure prediction. Some methods reduce the time complexity of the evaluation of internal loops from $O(n^4)$ to $O(n^3)$, reducing the overall complexity of secondary structure prediction to $O(n^3)$ (refer to [SAN1994, ZUK1984]. The partition function algorithm of [MCC1990] also calculates base pair probabilities in the thermodynamic ensemble and the suboptimal folding algorithm [WUC1999] generates all suboptimal structures within a given energy range of the optimal energy. For secondary structure comparison several measures of distances using either string alignment or tree-editing [SHA1990] have been defined.

- *Stochastic Context Free Languages*. The basic idea behind stochastic grammars is a direct extension of the HMMs [KNU1999]. Grammars are natural tools for modelling strings of letters and, more recently, they have been applied to biological sequences. In fact, many problems in computational molecular biology can be cast in terms of formal languages [SEA1997]. The basic goal again is to produce, by machine learning, the corresponding grammars from the data. In [EDD1994] an algorithm is reported by which the production rules themselves are derived from a set of unaligned sequences. For large RNA molecules, the process of constructing the grammar can also be hierarchically decomposed, whereby a high-level grammar [SAK1994] is first constructed on the basis of secondary-structure large-scale motifs [STE1993], such as helices and loops. Each motif is then separately represented by a set of corresponding grammar rules.

- *Sequence alignment and probability matrices.* Prediction of conserved secondary structure of a set of homologous single-stranded RNAs. For each RNA, the set the structure distribution is calculated and stored in a base pair probability matrix. These

'aligned' probability matrices are summed up to give a consensus probability matrix emphasizing the conserved structural elements of the RNA set. From the consensus probability matrix a consensus structure is extracted, which is viewable in three different graphical representations [LUC1999].

- *RNA tertiary structure prediction.* As the database of experimentally determined 3D structures containing RNA has only increased substantially in the recent years, the dominant method for computational prediction of these structures are molecular dynamics simulations with their known limitations [HU2003]. A promising approach for larger structures is the discrete conformational sampling using a database of known small interaction motifs as in the Mc-Sym system [LEM2002, GEN2001].

12.2.1 Available Systems and Tools

RNA secondary strucutre prediction

Vienna Package: http://www.tbi.univie.ac.at/~ivo/RNA/

Zuker: http://www.bioinfo.rpi.edu/~zukerm/

RNA tertiary structure prediction

NNPREDICT: http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html

Libellula: http://www.pdg.cnb.uam.es:8081/libellula.html

Mc-Sym: http://www-lbit.iro.umontreal.ca/mcsym/index.html

3DNA: http://rutchem.rutgers.edu/~olson/3DNA

12.2.2 References

[BEV2000] Beveridge, D. and McConnell, K. Nucleic acids: theory and computer simulation, Y2K. *Current Opinion in Structural Biology*, 10:182-196, 2000.

[EDD1994] Eddy, S.R. *et al.* RNA sequence analysis using covariance models. *Nucl. Acids Res.*, 22:2079-2088, 1994.

[FOG2002] Fogel, G. *et al.* Discovery of RNA structural elements using evolutionary computation. *Nucleic Acids Research*, 30(23):5310-5317, 2002.

[GEN2001] Gendron, P. *et al.* Quantitative Analysis of Nucleic Acid Three-dimensional Structures, J. Mol. Biol., 308:919-936, 2001.

[HU2003] Hu, H. *et al.* Comparison of a QM/MM Force Field and Molecular Mechanics Force Fields in Simulations of Alanine and Glycine "Dipeptides" (Ace-Ala-Nme and Ace-Gly-Nme) in Water in Relation to the Problem of Modeling the Unfolded Peptide Backbone in Solution. PROTEINS: Structure, Function, and Genetics, 50:451–463, 2003.

[JUA1999] Juan, V., Wilson, C. RNA secondary structure prediction based on free energy and phylogenetic analysis. J. Mol. Biol., 289:935-947, 1999.

[KNU1999] Knudsen, B., and Hein, J. RNA secondary structure prediction using stochastic context-free grammers and evolutionary history. *Bioinformatics*, 15(6):446-454, 1999

[LEM2002] Lemieux, S. and Major, F. RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire, Nucleic Acids Research, 30:4250-4263, 2002.

[LUC1999] Luck, R. *et al.* ConStruct: A tool for thermodynamic controlled prediction of conserved secondary structure. Nucleic Acids Res. 21, 4208-4217, 1999.

[MAJ2001] Major, F. and Griffey, R. Computational methods for RNA structure determination. Current Opinion in Structural Biology 2000, 11:282-286, 2001.

[MCC1990] McCaskill, J.S. The equilibrium partition function and base pair binding probabilities

for RNA secondary structure. *Biopolymers*, 29:1105-19, 1990.

[MEI2003] Meiler, J., and Baker, D. Coupled prediction of protein secondary and tertiary structure. *PNAS*, 100(21): 12105-12110, 2003.

[NIS1986] Nishikawa, K., Ooi, T. (1986) Amino-acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods. *Biochim Biophys Acta*, 871: 45-54.

[SAK1994] Sakakibara, Y., *et al.* D. Stochastic context-free grammars for modeling RNA. In Proceedings of the *27th Annual Hawaii International Conference on System Sciences.* Volume 5: Biotechnology Computing, L. Hunter, Ed. Los Alamitos, CA, USA: IEEE Computer Society Press, pp. 284—294, 1994.

[SAN1983] Sankoff, D. *et al.* Fast algorithms to determine RNA secondary structures containing multiple loops. In *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*, pp. 93-120, Sankoff D., Kruskal J.B., Eds. Addison-Wesley, Reading, MA, 1983.

[SEA1997] Searls, D.B. Linguistics approaches to biological sequences. *CABIOS*, 13:333-344, 1997.

[SHA1990] Shapiro, B.A. and Zhang, K.: Comparing multiple RNA secondary structures using tree comparisons. Comput. Appl. Biosci. 6: 309-318, 1990.

[STE1993] Steinberg, S. *et al.* Compilation of tRNA sequences and sequences of tRNA genes. *Nucl. Acids Res.*, 21:3011-3015, 1993.

[TIN1971] Tinoco, I.Jr. *et al.* Estimation of secondary structure in ribonucleic acids. Nature 230, 363-367, 1971.

[WUC1999] Wuchty, S. *et al.* Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145-65, 1999.

[ZUK1999] Zuker, M. *et al.* Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: RNA Biochemistry and Biotechnology (Ed.: Barciszewski *et al.*), 11-43, 1999.

[ZUK1986] Zuker, M. RNA folding prediction: The continued need for interaction between biologists and mathematicians. *Lectures on Mathematics in the Life Sciences*, 17:86-123, 1986.

[ZUK1984] Zuker, M., and Sankoff, D. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46:591-621, 1984.

[ZUK1981] Zuker, M., and Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9:133-148, 1981.

12.3 Protein structure prediction

The function of a protein is strongly dependent on its three-dimensional structure and one of the main challenges in bioinformatics is to be able to predict the structure of a protein from its primary sequence. The prediction of the native structure of a protein from its amino acid sequence remains an outstanding unsolved problem. In the post genomic era where protein structure can assist functional annotation, the need for progress is even more crucial. For proteins displaying significant sequence similarity to homologues with known structure, homology modelling has provided an effective tool for structure prediction. For the other cases, several methods have been proposed and developed to predict lower-level, usually discrete, protein structural features. The identification of these one-dimensional features represents progressive steps toward the prediction of the three-dimensional structure of proteins [BAL1999, ROS2001, SCH2002 and ZHA2004].

- *Protein secondary structure prediction.* Secondary structure predictions are increasingly becoming the work horse for numerous methods aimed at predicting protein structure and function. The evolutionary information resulting from improved searches and larger databases has again boosted prediction accuracy by more than four percentage points to

its current height of around 76% of all residues predicted correctly in one of the three states, helix, strand, and other. Utilized techniques include: (a) Neural networks - a method that is capable to use more sequence context for annotating sequences without increasing the number of adjustable parameters are the Bidirectional Recurrent Neural Networks (BRNN) that capture sequence context to the left and right of any position in a sequence in separate feedback loops – other ANN architectures are also utilized [MCG2000, POL2002, RAG2000]; (b) Hidden Markov Models – the approaches are similar to the ones mentioned for RNA secondary structure prediction [BYS2000] and (c) Support Vector Machines (SVM) – where, binary SVMs are trained to discriminate between two structural classes. The binary classifiers are combined in several ways to predict multi-class secondary structure [KIM2003, WAR2003].

- *Protein surface exposure prediction.* The prediction of protein relative solvent accessibility gives us helpful information for the prediction of tertiary structure of a protein and helps in identifying potential binding sites. The SVMpsi method, which uses support vector machines (SVMs), and the position-specific scoring matrix (PSSM) generated from PSI-BLAST have been applied to achieve better prediction accuracy of the relative solvent accessibility [KIM2004].

- *Protein disulfide bonding prediction using: (a) Hidden Markov Model/ Neural Network Hybrids.* The predictor accuracy is 88% after a 20-fold cross-validation procedure. Further, when tested on a protein basis, the hybrid system can correctly predict 84% of the chains in the data set, with a gain of at least 27% over the NN predictor [MAR2002]. (b) *Integer linear programming* has been applied to predict β architectures in polypeptides [KLE2003]. The approach is shown to perform very well for several benchmark polypeptide systems, as well as polypeptides exhibiting challenging non-sequential β sheet topologies folds (56 to 187 amino acids). (c) *Multiple sequence alignment.* This prediction classifies 82% of the cysteines in a jack-knife correctly [FIS2000]. (d) *Graph theory.* The problem of predicting the disulfide connectivity in proteins is equated to a problem of finding the graph matching with the maximum weight. In the case of proteins with four disulfide bonds in the structure the accuracy is 17 times higher than that of a random predictor [FAR2001].

- *Protein transmembrane region prediction.* Methods that predict membrane helices have become increasingly useful in the context of analyzing entire proteomes. Structurally unsolved multispanning membrane proteins, which are often important drug targets, will remain problematic for transmembrane helix prediction algorithms. The basic methods is use include: (a) *Artificial Neural Networks.* PHDhtm method combines a neural network using evolutionary information with a dynamic programming optimization of the final prediction. Tmbeta-net is a prediction of transmembrane beta-strands in outer membrane proteins (OMP). In this work, a method based on neural networks for identifying the membrane-spanning beta-strands is used. Predicted segments show a good agreement with experimental observations with an accuracy level of 73% solely from amino acid sequence information [GRO2004]. (b) *Hidden Markov Models.* TMHMM is the most advanced, and seemingly most accurate, present method to predict membrane helices [LIA2001, SON1998]. It embeds a number of statistical preferences and rules into a hidden Markov model to optimize the prediction of the localization of membrane helices and their orientation (similar concepts are used for HMMTOP) [TUS1998].

- *Protein tertiary structure prediction.* The quality of a structure model depends on how much information from already known structures can be used. The basic methodologies include: (a) Comparative or homology modelling: An approximate model can be created simply by copying related regions of polypeptide from the parent structures and changing the sidechains where necessary. TASSER is a system that follows this strategy and

implements a hierarchical approach to protein structure prediction that consists of template identification by threading, followed by tertiary structure assembly via the rearrangement of continuous template fragments and side-chain-based potential driven by threading-based, predicted tertiary restraints [GOU2001, SCH2003 and ZHA2004]. (b) Fold recognition: Increasingly, new structures deposited in the protein data bank turn out to have folds that have been seen before, even though there is no obvious sequence relationship between the related structures. Thus, methods of identifying folds from sequence information continue to grow in importance. ROSETTA Fragment assembly follows this approach. The Rosetta method of de novo protein structure prediction is based on the assumption that the distribution of conformations available to each three- and nine-residue segment of the chain is reasonably well approximated by the distribution of structures adopted by the sequence of the segment (and closely related sequences) in known protein structures. Fragment libraries for each three-and nine-residue segment of the chain are extracted from the protein structure database using a sequence and secondary structure profile- profile comparison method. The conformational space spanned by these fragments is then searched using a Monte Carlo procedure with an energy function that favours hydrophobic burial and strand pairing and disfavours steric clashes. For each target sequence, large numbers of decoy structures are generated with this protocol and then clustered; the five largest clusters are generally chosen as predictions [BRA2003].

- *Protein structural domain, functional sites and function prediction.* As the need for gene function prediction is so dominant, especially for interpreting the huge amounts of clinical genetics data, a shortcut predicting the protein domains [MAR2002] and their function [JEN2003] and avoiding the difficulties of structure prediction has been established recently. Additionally effective tools for predicting and classifying unstructured nonglobular functional motifs complementary to domain information help in direct functional annotation [PUN2003].

12.3.1 Available Systems and Tools

Protein secondary structure prediction

ANN/BRNN: http://www.igb.uci.edu/tools/scratch/CASP4_abstract.html

SVM/ESLPred: http://www.imtech.res.in/raghava/eslpred/

AGADIR: http://www.embl-heidelberg.de/Services/serrano/agadir/agadir-start.html

Coiled-coils: http://www.russell.embl.de/cgi-bin/coils-svr.pl

APSSP2: http://imtech.res.in/raghava/apssp/

GOR IV: http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html

PSIPRED: http://bioinf.cs.ucl.ac.uk/psipred/

Protein surface exposure prediction

SVMpsi (FORCAP server): http://www.forcasp.org/

PROSPECT: http://compbio.ornl.gov/structure/prospect2/

Protein disulfide bonding prediction

HMM: http://www.biocomp.unibo.it/

NYSGRC: http://www.nysgrc.org/

PROSPECT: http://compbio.ornl.gov/structure/prospect2/

MODELLER: http://www.salilab.org/modeller/methenz/index.html

Protein transmembrane region prediction

ANN/PHDhtm:http://www.embl-heidelberg.de/predictprotein/doc/methodsPP.html#PX_about_phdhtm

PredictProtein: http://cubic.bioc.columbia.edu/predictprotein/)

TMHMM: http://www.cbs.dtu.dk/services/TMHMM/

HMMTOP: http://www.enzim.hu/hmmtop/

TMBETA-NET: http://psfs.cbrc.jp/tmbeta-net/

Comparative modeling (homology / profiling)

SWISS-MODEL: http://swissmodel.expasy.org/

3Djigsaw: http://www.bmm.icnet.uk/servers/3djigsaw/

CPHmodels: http://www.cbs.dtu.dk/services/CPHmodels/

ESyPred3D: http://www.fundp.ac.be/urbm/bioinfo/esypred/

*Geno3d:*http://geno3d-pbil.ibcp.fr/cgi-bin/geno3d_automat.pl?page=/GENO3D/geno3d_home.html

SDSC1: http://cl.sdsc.edu/hm.html

Threading

3D-PSSM: http://www.sbg.bio.ic.ac.uk/~3dpssm/

Fugue: http://www-cryst.bioc.cam.ac.uk/~fugue/

LOOPP: http://ser-loopp.tc.cornell.edu/cbsu/loopp.htm

Threader: http://bioinf.cs.ucl.ac.uk/threader/threader.html

Protein tertiary structure prediction

TASSER: http://www.bioinformatics.buffalo.edu/new_buffalo/people/abinitio/1489/

ROSETTA Fragment assembly: http://graylab.jhu.edu/docking/rosetta/

PDB (Protein Data Bank): http://www.rcsb.org/pdb/

DALI: http://www.ebi.ac.uk/dali/; http://www.bioinfo.biocenter.helsinki.fi:8080/dali/index.html

MMDB: http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml

VAST: http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html

SCOP: http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/,

SWISS-Model: http://www.expasy.org/swissmod/SWISS-MODEL.html

Protein domain and function prediction

ELM (Eukaryotic Linear Motif) server: http://elm.eu.org

ProtFun 2.2 Server: http://www.cbs.dtu.dk/services/ProtFun

12.3.2 References

[BAL1999] Baldi, P. *et al.* Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics,* 15:937–946, 1999.

[BER2000] Berman, H.M. *et al.* The Protein Data Bank. *Nucleic Acids Research*, 28:235-242, 2000.

[BRA2003] Bradley, P. et. al. Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation. *PROTEINS*, 53:457-468, 2003.

[BYS2000] Bystroff, C. *et al.* HMMSTR: A hidden Markov model for local sequence–structure correlations in proteins, *J. Mol. Biol., 301:173*–190, 2000.

[FAR2001] Fariselli, P. *et al.* Prediction of the disulfide connectivity in proteins. *Bioinformatics*, 17(10):957-964, 2001.

[FIS2000] Fiser, A., Simon, I. Predicting the oxidation state of cysteines by multiple sequence alignment. *Bioinformatics*, 16(3):251-256, 2000.

[GOU2001] Gough, J. *et al.* Assignment of Homology to Genome Sequences using a Library of Hidden Markov Models that Represent all Proteins of Known Structure. *J. Mol. Biol.*, 313(4):903-919, 2001.

[GRO2004] Gromiha, M. *et al.* Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins. *J. of Computational Chemistry*, 25:762-767, 2004.

[HOL1996] Holm, L., and Sander, C. Mapping the protein universe. *Science*, 273:595-602, 1996.

[JEN2003] Jensen, L. *et al.* Prediction of human protein function according to Gene Ontology categories. *Bioinformatics*, 19:635-642, 2003.

[KIM2003] Kim, H., and Park, H. Protein secondary structure prediction based on an improved support vector machines approach. *Protein Engineering*, 16(8):553-560, 2003.

[KIM2004] Kim, H., Park, H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. *PROTEINS: Structure Function and Genettics*, 54:557-562, 2004.

[KLE2003] Klepeis J.L. and C.A. Floudas, "Prediction of Beta-Sheet Topology and Disulfide Bridges in Polypeptides", Journal of Computational Chemistry, 24:191-208, 2003.

[LIA2001] Liakopoulos, T. *et al.* A novel tool for the prediction of transmembrane protein topology based on a statistical analysis of the SwissProt database: the OrienTM algorithm. *Proteinn Engineering*, 14(6):387-390, 2001.

[MAR2002a] Marsden, R. *et al.* Rapid protein domain assignment from amino acid sequence using predicted secondary struc ture Protein. *Science* 11:2814–2824, 2002.

[MAR2002b] Martelli, P. *et al.* Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks. *Protein Engineering*, 15(12):951–953, 2002.

[MCG2000] McGuffin L. *et al.* The PSIPRED protein structure prediction server. *Bioinformatics* 16, 404-405, 2000.

[POL2002] Pollastri, G. *et al.* Improving the prediction of protein secondarystructure in three and eight classes using recurrent neural networks andprofiles. *Proteins: Structure, Function, and Genetics*, 47:228-235, 2002.

[PUN2003] Puntervoll, P. *et al.* ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Research* 31(13):3625–3630,

2003.

[RAG2000] Raghava, G. P. S. Protein secondary structure prediction using nearest neighbor and neural network approach. CASP4: 75-76, 2000.

[ROS2001] Rost, B. Review: Protein secondary structure prediction continues to rise. *J. of Structural Biology*, 134:204-218, 2001.

[SCH2002] Schonbrun, J. *et al.* Protein structure prediction in 2002. *Current Opinion in Structural Biology*, 12:348-354, 2002.

[SCH2003] Schwede, T. *et al.* SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Research*, 31:3381-3385, 2003.

[SON1998] Sonnhammer, E. *et al.* A hidden Markov model for predicting transmembrane helices in protein sequences. In *Sixt h International Conference on Intelligent Systems for Molecular Biology (ISMB98)* (eds. J. Glasgow), pp. 175–182. AAAI Press, Montreal, Canada, 1998.

[TUS1998] Tusnady, G.E., Simon, I. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol.*, 283(2):489-506, 1998.

[WAR2003] Ward, J. *et al.* Secondary structure prediction with support vector machines. *Bioinformatics*, 19:1650-1655, 2003.

[ZHA2004] Zhang. Y., and Skolnick,.J. Automated structure prediction of weakly homologous proteins on a genomic scale. *PNAS*, 101:7594-9, 2004.

12.4 Bio-Molecular Interaction and Pathway Modelling

Interest in protein-protein interaction has grown fast during the last few years, largely as the outcome of proteomics studies such as genome wide yeast two-hybrid assays or high-throughput mass spectrometry [BAD2002]. These studies show that most if not all proteins have interacting partners in the cell. A major goal of functional genomics is to determine protein interaction networks for whole organisms [BOC2003]. Experimental methods that can a globally tackle the problem have been developed [UET2000],[GAV2002]. Those high-throughput methods have led to the creation databases containing large sets of protein interactions, such as DIP [SAL2004], MIPS [MEW2004] and HPRD [PER2004].

Several in silico methods have been developed to predict protein-protein interactions based on gene context features. These include gene fusion, [MAR1999] gene neighbourhood [DAN199] and phylogenetic profiles [Pellegrini *et al.* 1999]. The emerging map of protein-protein interactions is a major challenge to experimentalists and also to computational biologists.

An emerging new approach in the protein interactions field is to take advantage of structural information to predict physical binding [ALO2004]. Several decades of X-ray crystallography have produced hundreds of structures for protein complexes. Moreover, recent technical advances have allowed the application of NMR to large protein complexes [FIA2002] . Although the total number of number of complexes of known structure is relatively small, it is possible to expand this set by considering homologous proteins. The majority of cases of close homologues (above 30% sequence identity) physically interact in the same way with each other [RUS2004] . However, conservation of a particular interaction depends on the conservation of the interface between interacting partners. Thus, computational methods that can elucidate the details of specific protein-protein interaction at the atomic level are becoming of greater value as more structures of individual proteins are determined while the protein-protein interaction map expands. Resolving one important interaction is assisted by water binding site prediction [EHR1998]. The ability to model the docking of two proteins is fundamental to the understanding of the operation of biochemical systems. The basic

methodologies utilised are: (a) Genetic algorithms which generate rotations of the smaller protein relative to the larger protein surface, which is held static [GAR2003]; and (b) Threading approaches; underlying tools include, COBLATH - a structure-prediction method that exploits the complementarity of PSI-Blast and sequence-structure threading [ZHO2004]; and PPISP - a method that predicts the residues involved *in protein-protein interactions* [http://www.csit.fsu.edu/ ~hxzhou/reprints/pr50.pdf].

12.4.1 Available Systems and Tools

COBLATH: http://cmbph4.physics.drexel.edu/COBLATH/
PPISP: http://cmbph1.physics.drexel.edu/cgi-bin/PPISP.cgi
DIP: http://dip.doe-mbi.ucla.edu/
MIPS: http://mips.gsf.de/
HPRD: http://www.hprd.org/
STRING: http://string.embl.de/
InterPRETS: http://www.russell.embl.de/interprets/

12.4.2 References

[ALO2004] Aloy, P., and Russell, R.B. 2004. Ten thousand interactions for the molecular biologist. Nat Biotechnol 22: 1317-1321.

[BAD202] Bader, G.D., Hogue, C.W. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol.*, 20(10):991-997, 2002.

[BOC203] Bock JR, Gough DA. Whole-proteome interaction mining. *Bioinformatics*, 19(1):125-134, 2003.

[DAN1998] Dandekar, T., Snel, B., Huynen, M., and Bork, P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 23: 324-328.

[DEA2002] Deane, C.M., Salwinski, L., Xenarios, I., Eisenberg, D. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics*, 1(5):349-356, 2002.

[EHR1998] Ehrlich, L., Reczko, M., Bohr, H. and Wade, R.C. Prediction of protein hydration sites from sequence by modular neural networks. *Protein Eng*, 11(1):11-19, 1998.

[FIA2002] Fiaux, J., Bertelsen, E.B., Horwich, A.L., and Wuthrich, K. 2002. NMR analysis of a 900K GroEL GroES complex. Nature 418: 207-211.

[GAR2003] Gardiner, E. *et al.* GAPDOCK: A genetic algorithm approach to protein docking in CAPRI round 1. *PROTEINS: Structure, Function, and Genetics*, 52:10-14, 2003.

[GAV2002] Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., *et al.* 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415: 141-147.

[HER2004] Hermjakob, H., et. al. The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. *Nat. Biotechnol.*, 22(2):177-83, 2004.

[IOS2004] Iossifov, I., Krauthammer, M., Friedman, C., Hatzivassiloglou, V., Bader, J.S., White, K.P., Rzhetsky, A. Probabilistic inference of molecular networks from noisy data sources. *Bioinformatics*, 10, 2004.

[MAR1999] Marcotte, E., Pellegrini, M., Ho-Leung, Rice, D.W., Yeates, T.O., and Eisenberg, D. 1999. Detecting Protein Function and protein-protein interactions from genome sequences. Science 285: 751-753.

[MEW2004] Mewes, H.W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., *et al.* 2004. MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res 32 Database issue: D41-44.

[NG2003] Ng, S.K., Zhang, Z., Tan, S.H. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19(8):923-929, 2003.

[PEL1999] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285-4288.

[PER2004] Peri, S., Navarro, J.D., Kristiansen, T.Z., Amanchy, R., Surendranath, V., Muthusamy, B., Gandhi, T.K., Chandrika, K.N., Deshpande, N., Suresh, S., *et al.* 2004. Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res 32 Database issue: D497-501.

[RUS2004] Russell, R.B., Alber, F., Aloy, P., Davis, F.P., Korkin, D., Pichaud, M., Topf, M., and Sali, A. 2004. A structural perspective on protein-protein interactions. Curr Opin Struct Biol 14: 313-324.

[SAL2004] Salwinski, L., Miller, C., Smith, A., Pettit, F., Bowie, J., and Eisenberg, D. 2004. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. 32: D449-451.

[TAN2004] Tanay, A., Sharan, R., Kupiec, M., and Shamir, R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *PNAS*, 101(9): 2981-2986, 2004.

[TON2002] Tong, A.H. *et al.* A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, 295(5553):321-4, 2002.

[UET2000] Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., *et al.* 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403: 623 - 627.

[VON2002] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., Bork, P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 23:417(6887):399-403, 2002.

[ZHO2004] Zhou, H. Improving the understanding of human genetic diseases through predictions of protein structures and protein-protein interaction sites. Current Medicinal Chemistry, 11(5):539-549, 2004.

12.5 Microarrays and Gene Expression Profiling

Microarray data analysis is heavily dependent on Gene Expression Data Mining (GEDM) technology, and in the very-last years a lot of research efforts are in progress. GEDM is used to identify intrinsic patterns and relationships in gene expression data. The identification of patterns in complex gene expression datasets provides two benefits:

- Generation of insight into gene transcription conditions.
- Characterization of multiple gene expression profiles in complex biological processes, e.g. pathological states.

GEDM activities are based on two approaches: (a) *Hypothesis testing*: to investigate the induction or perturbation of a biological process that leads to predicted results, and (b) *Knowledge Discovery*: to detect internal structure in biological data. References, of general flavour about gene expression experiments and related tasks, are listed below.

12.5.1 Available Systems and Tools

Affymetrix: http:/www.affymentrix.com

Qiagen: http://www.giagen.cpm/

SAGE: http://www.sagnet.org/

12.5.2 References

[DUT2002] Dutton, G. Gene Expression Data Mining. The Scientist, 16(20), 2002.

[GLO2004] Glonek, G.F., and Solomon, P.J. Factorial and time course designs for cDNA microarray experiments. *Biostatistics*, 5(1):89-111, 2004.

[HWA2002] Hwang, D., Schmitt, W.A., Stephanopoulos, G, and Stephanopoulos, G. Determination of minimum sample size and discriminatory expression patterns in microarray data. *Bioinformatics*, 18(9):1184-93, 2002.

[KER2003] Kerr, M.K. Experimental design to make the most of microarray studies. *Methods Mol Biol.*, 224:137-47, 2003.

[KER2001] Kerr, M.K., Churchill, G.A. Experimental design for gene expression microarrays. *Biostatistics*, 2(2):183-201, 2001.

[LEE2002] Lee, M.L., and Whitmore, G.A. Power and sample size for DNA microarray studies. *Stat Med.*, 21(23):3543-70, 2002.

[SIM2002] Simon, R., Radmacher, M.D., and Dobbin, K. Design of studies using DNA microarrays. *Genet Epidemiol.* 23(1):21-36, 2002.

[SIM2003] Simon, R.M., and Dobbin, K. Experimental design of DNA microarray experiments. Biotechniques, Mar; Suppl: 16-21, 2003.

12.6 Intelligent Processing of Gene-Expression Data

By measuring transcription levels of genes in an organism under various conditions or in different tissues we can build up 'gene expression profiles', which characterize the dynamic function of each gene. Microarray data are represented in a matrix with rows representing genes, columns representing samples and each cell containing a number characterizing the gene expression level in the particular sample, i.e., the gene expression matrix. *I*ndicative references about microarrays and gene expression profiling methodologies are included below.

GENE EXPRESSION DATA-MINING SUPPLIERS

COMPANY	WEBSITE	PRODUCT
Affymetrix	www.affymetrix.com	Netaffx™ Analysis Center, Data Mining Tool
Axon Instruments	www.axon.com	Acuity™
BioDiscovery	www.biodiscovery.com	Gene Director™ GeneSight DB™
Biomax Informatics	www.biomax.de	BioRS, Pedant-Pro, HarvESTer
Compugen	www.cgen.com	GeneGuide
Gene Network Sciences	www.gnsbiotech.com	BioMine™
IBM	www.ibm.com	GeneMine™
InforMax	www.informaxinc.com	GenoMax, Xpression NTI
Incyte Genomics	www.incyte.com	ChemExpress™, SNooPer,
Insightful	www.insightful.com	InFact®
Iobion Informatics	www.iobion.com	Gene Traffic™
LION bioscience	www.lionbioscience.com	DiscoveryCenter™, arraySCOUT™
MiraiBio	www.miraibio.com	DNAISIS
Molecular Mining	www.molecularmining.com	GeneLinker™ Platinum
Rosetta Inpharmatics	www.rii.com	Rosetta Resolver
SAS Institute	www.SAS.com	Enterprise Miner
Scimagix	www.scimagix.com	Scientific Image Management System,
Silicon Genetics	www.sigenetics.com	MetaMine
Silicon Graphics	www.sgi.com	MineSet
Spotfire	www.spotfire.com	DecisionSite for Functional Genomics
SPSS	www.spss.com	Clementine
Visual Bioinformatics	www.visual-bioinformatics.com	GeneWeaver™
X-MINE	www.x-mine.com	X-Miner

12.6.1 References

[BAS1999] Bassett, D.E., Eisen, M.B. and Boguski, M.S. Gene expression informatics: it's all in your mine. *Nat. Genet.* 21(Supl. 1):51-55, 1999.

[BRA2000] Brazma, A., and Vilo, J. Gene expression data analysis. *FEBS Lett.* 480(1):17-24, 2000.

[LEU2003] Leung, Y.F., Cavalieri, D. Fundamentals of cDNA microarray data analysis. *Trends Genet.* 19(11):649-59, 2003.

[NAD2002] Nadon, R., Shoemaker, J. Statistical issues with microarrays: processing and analysis. *Trends Genet.* 18(5):265-71, 2002.

[QUA2001] Quackenbush, J. Computational genetics computational analysis of microarray data. *Nat Rev Genet.*, 2(6):418-27, 2001.

[SCH1998] Schena, M., Heller, R.A., Theriault, T.P., Konrad, K., Lachenmeier, E. and Davis, R.W. Microassays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol.* 16(7):301-306, 1998.

[SHE2001] Sherlock, G. Analysis of large-scale gene expression data. *Brief Bioinform.* 2(4):350-62, 2001.

[SMY2003] Smyth GK, Yang YH, Speed T. Statistical issues in cDNA microarray data analysis. *Methods Mol Biol.*, 224:111-36, 2003.

[SOR2001] Sorlie, T., Perou, C., Tibshiranie, R., Aasf, T., Geislerg, S., Johnsen, H., Hastiee, T., Eisen, M., Van de Rijni, M., Jeffreyj, S., Thorsenk, T., Quistl, H., Matesec, J., Brown, P., Botstein, D., Lonningg, P.E., and Borresen-Dale, A., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. PNAS*, 98(19):10869-10874, 1001.

[STR1999] Stratowa, C. and Wilgenbus, K.K. (1999). Gene expression profiling in drug discovery and development. *Curr. Opin. Mol. Ther.,* 1(6):671-679, 1999.

[WU2001] Wu, T.D. Analysing gene expression data from DNA microarrays to identify candidate genes. *J Pathol*. 5(1):53-65, 2001.

[ZWE1999] Zweiger, G. Knowledge discovery in gene-expression-microarray data: mining the information output of the genome. *Trends Biotechnol.* 17(11):429-436, 1999.

12.6.2 Microarrays and Image Analysis

Over the last years there has been a substantial effort to develop techniques for the effective measurement and Imaging of Gene expression as an essential step:

- for enabling systematic progress in clinical and genetic data interoperability and integration.
- for advancing the exchange and interfacing of methods, tools and technologies used in both Medical Informatics and Bioinformatics.

To date, the main techniques used for the measurement and Imaging of Gene expression are:

- Microarray Imaging: This Imaging technique is based on an array of DNA or protein samples that can be hybridized with probes to study patterns of gene expression.

- Multiplex or multi-colour fluorescence in situ hybridization (M-FISH) imaging: By simultaneously viewing the multiple-labelled specimens in different colour channels, M-FISH facilitates the detection of subtle chromosomal aberrations.

The main problem of these techniques is the following: the intensity measured in each spot includes a contribution of non-specific hybridization and other chemicals on the glass; this is probably one of the most significant and profound problems in microarray data analysis since the influence of the background and the scanner are not completely understood. As a result, the task of discriminating healthy from diseased tissue becomes problematic since we can't directly relate gene expression to the measured intensity. Although there have been proposed methods to estimate the background locally, the issue of normalization remains an open research topic.

More detailed presentation of microarray image processing problems, techniques, systems, methods and tools is given in a following section.

12.6.2.1 Available Systems and Tools

Academic (free)

BASE: http://base.thep.lu.se/
Dapple: http://www.cs.wustl.edu/%7Ejbuhler//research/dapple/
F-SCAN: http://abs.cit.nih.gov/fscan/
GeneSpotter: http://www.microdiscovery.de/en/psGsMainEn.php
GrodGrinder: http://Gridgrinder.sourceforge.net/
MIDAS: http://www.tigr.org/software/tm4/midas.html
P-SCAN: http://abs.cit.nih.gov/pscan/index.html
ScanAlyze: http://rana.lbl.gov/downloads/ScanAlyze.zip *(Eisen's Lab)*
Spotfinder : http://www.tigr.org/software/tm4/spotfinder.html
UCSF Spot: http://www.jainlab.org/downloads.html

Commercial (demo/trial versions)

ArrayFox: http://www.imaxia.com/products.htm
Array-Pro Analyzer : http://www.mediacy.com/arraypro.htm
ArrayVision : http://www.imagingresearch.com/products/ARV.asp
GenePix : http://www.axon.com/GN_GenePixSoftware.html
IconoClust : http://www.clondiag.com/products/sw/iconoclust/
ImaGene : http://www.biodiscovery.com/imagene.asp
ImaGene: http://www.biocompare.com/itemdetails.asp?itemid=130725&catid=2975
Koadarray: http://www.koada.com/koadarray/
MacroView: http://www.genefocus.com/specifications_macroview_software.htm
MicroVigene: http://www.vigenetech.com/product.htm
Phoretix Array: http://www.nonlinear.com/products/array/
SilicoCyte: http://www.silicocyte.com/dis/imageanalysis.htm
Spot : http://experimental.act.cmis.csiro.au/Spot/index.php
SpotReader: http://www.nilesscientific.com/welcome.nhtml

12.6.2.2 References

[ANG2003] Angulo, J. and Serra, J. Automatic analysis of DNA microarray images using mathematical morphology. *Bioinformatics*, 19(5):553-562, 2003.

[DES2002] De Smet, F., Mathys, J., Marchal, K., Gert Thijs, G., De Moor, B., and Moreau, Y. Adaptive Quality-based clustering of gene expression profiles. *Bioinformatics*, 18(6):735-746, 2002.

[FIL2002] Filkov, V, Skiena, S., and Zhi, J. Analysis Techniques for Microarray Time-Series Data, *J. of Computational Biology*, 9(2):317–330, 2002.

[QUA2001] Quackenbush, J. Computational Analysis of Microarray data. *Nature Reviews Genetics*, 2:418-427. 2001.

[SAA2002] Saal, L.H., Troein, C., Vallon-Christersson, J., Gruvberger, S., Borg, A., and Peterson, C. BioArray Software Environment: A Platform for Comprehensive Management and Analysis of Microarray Data. *Genome Biology*, 3(8), 2002.

[TUR2004] Turkheimer, F.E., Duke, D.C., Moran, L.B., and Graeber, M.B. Wavelet analysis of Gene expression (WAGE). *IEEE International Symposium on Biomedical Imaging*, Arlington VA, 2004.

[VAC2004] Vachtsevanos, G., Ding, Y., Fairley, J.A., Gardner, A.B., and Simeonova, P.Microarray Gene expression data analysis. *IEEE International Symposium on Biomedical Imaging*, Arlington VA, 2004

[WAN2003] Wang, X.H., Istepanian, R.S.H., and Song, Y.H. Microarray Image Enhancement by Denoising Using Stationary Wavelet Transform, *IEEE Trans on Nanobioscience*, 2(4), 2003.

[WAN2004] Wang, Y-P. M-FISH image registration and classification. *IEEE International Symposium on Biomedical Imaging*, Arlington VA, 2004.

[YAN2002] Yang, Y.H., Dudoit, S, Luu, P., Lin, D.M., Peng, V., Ngai, J., and Speed, T.P. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Res*., 30(4), 2002.

12.6.3 Microarray Data Pre-Processing and Normalization

Because there are many sources of noise and systematic variability in microarray experiments [SCH2000; WIL2001] data normalization and pre-processing are crucial in analysis [SCH2000]. Normalization includes those transformations that control systematic variabilities within a chip or across multiple chips. The simplest way data normalization can be done is by dividing or subtracting all expression values by a representative value for the system or by a linear transformation to a fixed mean (i.e., 0.0) and unit variance (i.e., 1.0) (sometimes called “median polishing”). However, the linear response between the true expression level and measured fluorescent intensity may not be guaranteed [KEP2002; TSE2001], especially when dye biases depend on array spot intensity or when multiple print tips are used in the microarray spotter [YAN2001; YAN2002].

Data pre-processing includes those transformations that prepare the data for the subsequent analysis. Scaling and filtering are the major steps of data pre-processing. A low variation filter to exclude genes that did not change significantly across experiments has been successfully applied in many studies [TAM1999]. Statistical significance testing, such as the analysis of variance and multiple comparisons, can also be used to filter data that show no significant change across conditions when a sufficient number of repeated observations are available. The importance of data visualization cannot be overemphasized. It is highly recommended to scatter-plot the data whenever possible. The most straightforward approach to microarray data analysis is to find differentially expressed genes across different experimental conditions [DER1996; HEL1997]. Standardized expression profiling, consistent database design, and streamlining the experimental process management are all crucial [BRA2001;

PER2001] as are the supervised and unsupervised machine-learning algorithms that make sense of the mountains of genomic data. Here now is a brief description of the various machine-learning approaches to deciphering genomic data.

12.6.3.1 Available Systems and Tools

Academic (free / refer also to 'microarray image analysis software')

ANOVA-based: http://genome1.beatson.gla.ac.uk/Rweb/anova.html

ArrayNorm: http://genome.tugraz.at/Software/GenesisCenter.html

GEPAS server: http://gepas.bioinfo.cnio.es/cgi-bin/preprocess

Microhelper: http://www.changbioscience.com/microhelperinfo.html

SNOMAD: http://pevsnerlab.kennedykrieger.org/snomadinput.html

TM4 (TIGR MIDAS): http://www.tigr.org/software/tm4/

Commercial (demo/trial versions)

Genemaths XT: http://www.applied-maths.com/genemaths/genemaths.htm

Predictive Patterns: http://www.predictivepatterns.com/products/index.html

12.6.3.2 References

[BRA2001] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. (2001). Minimum information about a microarray experiment (MIAME): toward standards for microarray data. *Nat Genet* 29, pp. 365–371.

[DER1996] DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM. Use of a cDNA microarray to analyse gene expression patterns in human cancer. (1996) *Nat Genet* 14, pp. 457–460.

[HEL1997] Heller RA, Schena M, Chai A, Shalon D, Bedilion T, Gilmore J, Woolley DE, Davis RW. (1997). Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci U S A* 94, pp. 2150–2155.

[KEP2002] Kepler TB, Crosby L, Morgan KT. (2002). Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biol* 3, RESEARCH0037.

[PER2001] Perou CM. Show me the data! (2001). *Nat Genet* 29:373.

[SCH2000] Schadt, E.E., Li, C., Su, C., and Wong, W.H. (2000). Analyzing high-density oligonucleotide gene expression array data. *J. Cellul. Biochem*. 80, pp. 192-202.

[SCH2000] Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, Lehrach H, Herzel H. Normalization strategies for cDNA microarrays. (2000). *Nucleic Acids Res* 28:10-E47.

[TAM1999] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. (1999). *Proc Natl Acad Sci U S A* 96, pp. 2907–2912.

[TSE2001] Tseng GC, Oh M, Rohlin L, Liao JC, and Wong WH. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variation and assessment of gene effects. (2001). *Nucleic Acids Res* 29, pp. 2549–2557.

[WIL2001] Wildsmith SE, Archer GE, Winkley AJ, Lane PW, Bugelski PJ. (2001). Maximization of signal derived from cDNA microarrays. *Biotechniques* 30, pp. 202–206, 208.

[YAN2001] Yang YH, Dudoit S, Luu P, Speed TP. (2001). Normalization for cDNA microarray data. SPIE BiOS 2001, San Jose, CA, January 2001.

[YAN2002] Yang, Y.-H., Dudoit, S., Lu, P., Lin, D.M., Peng, V., Ngai, J., and Speed, T.P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucl. Acids Res.* 30:4.

12.6.4 Clustering and Gene Expression Profiling

Cluster analysis is currently the most frequently used multivariate technique to analyze microarray data. Clusters can be developed using a variety of similarity or distance metrics: Euclidean distance, correlation coefficients, or mutual information.

- *Hierarchical* tree clustering joins similar objects together into successively larger clusters in a bottom-up manner (i.e., from the leaves to the root of the tree), by successively relaxing the threshold of joining objects or sets [EIS1998; IYE1999].

- The *relevance-networks* approach takes the opposite strategy [BUT2000]. It starts with a completely connected graph with the vertices representing each object and the edges representing a measure of association, and then links are increasingly deleted to reveal "naturally emerging" clusters at a certain threshold.

- Creation of a *hierarchical-tree* structure in a top-down fashion (i.e., from the root to the leaves of the tree) by successive "optimal" binary partitioning based on *graph theory* [SHA2000, POT2004; POT2004] and geometric space-partitioning principle [KIM2001] has also been introduced. The "optimal" partitioning problem (i.e., the best clustering) is fundamentally NP-hard and can be viewed as an optimization problem. Most of the meta-heuristic algorithms, such as simulated annealing and genetic algorithm [LEE2001] and model-based search [YEU2001] can all be applied to attain better understanding of the complex data structure of genomic-scale expression profiles.

- *Partitional* clustering algorithms, such as *K*-means analysis and self-organizing maps [KOH2982] which minimize within-cluster scatter or maximize between-cluster scatter, were shown to be capable of finding meaningful clusters from functional genomic data [TAV1982; TAM1999].

The reliability and quality measures of clusters, as well as multilevel visualization (see also section 2.5.2) for the evaluation of clustering solutions, should be addressed as well [YEU2001].

12.6.4.1 Available Systems and Tools

Academic (free)

AMIADA: http://aix1.uottawa.ca/~xxia/software/amiada.html

BRB Array Tools: http://linus.nci.nih.gov/BRB-ArrayTools.html

Cleaver: http://classify.stanford.edu/k-means.html

CLUSFAVOR: http://mbcr.bcm.tmc.edu/genepi/clusfavor.html

CLUSTER (Eisen Lab): http://rana.lbl.gov/EisenSoftware.htm

Cluster Control: http://genome.tugraz.at/Software/GenesisCenter.html

CTWC: http://ctwc.weizmann.ac.il/

dCHIP: http://www.dchip.org/

Expression Profiler: http://ep.ebi.ac.uk/EP/

Freeview: http://magix.fri.uni-lj.si/freeview/ (A java implementation with some functional enhancement of the famous *Treeview* program; accepting input from Cluster)

GEPAS server: http://gepas.bioinfo.cnio.es

INCLUsive: http://www.esat.kuleuven.ac.be/%7Edna/Biol/Software.html

MAExplorer: http://www.lecb.ncifcrf.gov/MAExplorer/

MCD: http://www.thep.lu.se/~markus/software/classdiscoverer/

SAM: http://www-stat.stanford.edu/%7Etibs/SAM/index.html (Excel add-in)

GeneClustre 2: http://www.broad.mit.edu/cancer/software/genecluster2/gc2.html

Commercial (demo/trial versions)

AQUITY: http://www.axon.com/GN_Acuity.html

ArrayMiner: http://www.optimaldesign.com/ArrayMiner/ArrayMiner.htm

Genemaths XT: http://www.applied-maths.com/genemaths/genemaths.htm

GeneSpring: http://www.silicongenetics.com/cgi/SiG.cgi/Products/GeneSpring/index.smf
http://www.partek.com/html/products/predict.html

Partek: http://www.partek.com/html/products/discover.html;

Rosetta Resolver: http://www.rosettabio.com/products/resolver/default.htm

Predictive Patterns: http://www.predictivepatterns.com/products/index.html

12.6.4.2 References

[BUT2000] Butte AJ, Kohane IS. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput, pp.* 418– 429.

[EIS1998] Eisen MB, Spellman PT, Brown PO, Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95, pp. 14863–14868.

[IYE1999] Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson J Jr, Boguski MS, Lashkari D, Shalon D, Botstein D, Brown PO. (1999). The transcriptional program in the response of human fibroblasts to serum. *Science* 283, pp. 83–87.

[KIM2001] Kim JH, Ohno-Machado L, Kohane IS. (2001). Unsupervised learning from complex data: the matrix incision tree algorithm. *Pac Symp Biocomput* pp. 30–41.

[KOH1982] Kohonen T. (1982). Self-organized formation of topologically correct feature maps. *Biol Cybern* 43, pp. 59–69.

[LEE2001] Lee K, Kim JH, Chung TS, Moon BS, Lee H, Kohane IS. (2001). Evolution strategy applied to global optimization of clusters in gene expression data of DNA microarrays. Proceedings of IEEE Congress on Evolutionary Computation, Seoul, Korea, May 27–30, 2001, pp. 845–850.

[POT2004a] Potamias G. (2004). Knowledgeable Clustering of Microarray Data. *LECT NOTES COMPUT SC – LNCS* 3337, pp. 491– 497.

[POT2004b] Potamias G., and Dermon C. (2004). Protein Synthesis Profiling in the Developing Brain: A Graph Theoretic Clustering Approach. *Computer Methods and Programs in Biomedicine*, 76, pp. 115– 129.

[SHA2000] Sharan R, Shamir R. (2000). CLICK: a clustering algorithm with applications to gene expression analysis. *Proc Int Conf Intell Syst Mol Biol* 8, pp. 307–316.

[TAM1999] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander E, Golub TR. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*

96, pp. 2907–2712.

[TAV1999] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. (1999). Systematic determination of genetic network architecture. *Nat Genet* 22, pp. 281–285.

[YEU2001] Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17, pp. 977–987.

12.6.5 Classification & Gene Expression Profiling

Classification is a supervised intelligent data analysis approach. One of the goals of supervised expression data analysis is to construct classifiers, such as decision trees or, rules; support vector machines (SVM), or, trained Artificial Neural Networks (ANN), which assign predefined classes to a given expression profile potentially to be used for diagnostics. By comparing samples, we can find classification-archetypes (class descriptions) with which differentially expressed genes are combined to distinguish between the samples and *'discriminant'* genes might be identified. For indicative references about microarrays and gene expression classification refer to the ULR links and references below.

12.6.5.1 Available Systems and Tools

Academic (free)

GEPAS server: http://gepas.bioinfo.cnio.es
AFFYR PACE: http://www.cbs.dtu.dk/staff/laurent/download/affyR/
GenePattern : http://www.broad.mit.edu/cancer/software/genepattern/
Boosting: http://stat.ethz.ch/~dettling/boosting.html
GeneClustre 2: http://www.broad.mit.edu/cancer/software/genecluster2/gc2.html

12.6.5.2 References

[BRO1999] Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C., Ares, M. Jr., and Haussler, D. Support Vector Machine Classification of Microarray Gene Expression Data. *University of California, Santa Cruz, Technical Report UCSC-CRL-99-09*.

[CAU2003] Causton, H.C, Quackenbush, J., and Brazma, A. *Microarray Gene Expression Data Analysis: A Beginner's Guide*. Blackwell Publishing, 2003.

[DET2003] Dettling, M., and Buhlmann, P. Boosting for tumor classification with gene expression data. Bioinformatics, 19(9):1061-1069, 2003.

[EIL2001] Eilers, P.H.C., Boer, J.M., van Ommen, G. –J., and van Houwelingen, H.C. Classification of microarray data with penalized logistic regression. *Proc. Int. Symp. Biomedical Optics*, San Jose, 2001.

[GHO2002] Ghosh, D. Singular value decomposition regression models for classication of tumors from microarray experiments. *Pacific Symposium on Biocomputing*, 7:18-29, 2002.

[GOL1999] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531-537, 1999.

[MIT1997] Mitchell, M. T. *Machine Learning*. McGraw-Hill, 1997.

[MUK1999] Mukherjee, S. *et al.* Support Vector Machine Classification of Microarray Data. *AI*

MEMO, CBCL Paper No. 182, 1999.

[NGU2002] Nguyen, D.V., and Rocke, D.M. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18:39-50, 2002.

[POT2004] Potamias G., Koumakis L., and Moustakis V. Gene Selection via Discretized Gene-Expression Profiles and Greedy Feature-Elimination. *LECT NOTES ARTIF INT (LNAI)*, 3025:256-266, 2004.

[POT2004] Potamias, G. Knoweldgable Clustering of Microarray Data. *International Symposium on Biological and Medical Data Analysis (ISBMDA-2004)*, Barcelona, Spain, 2004.

[QUI1993] Quinlan, J. R. *C4.5: Programs for Machine Learning*. San mateo, CA: Morgan Kaufmann. 1993.

[SLO2000] Slonim, D.K., Tamayo, P., Mesirov, J.P., Golub, T.R., and Lander, E.S. Class prediction and discovery using gene expression data. *Proceedings RECOMB IV*:263-271, 2000.

[SOR2001] Sorlie, T., Perou, C., Tibshiranie, R., Aasf, T., Geislerg, S., Johnsen, H., Hastiee, T., Eisen, M., Van de Rijni, M., Jeffreyj, S., Thorsenk, T., Quistl, H., Matesec, J., Brown, P., Botstein, D., Lonningg, P.E., and Borresen-Dale, A., Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *PNAS*, 98(19):10869-10874, 2001.

[SPE2003] Speed, T. *Statistical Analysis of Gene Expression Microarray Data*. Chapman and Hall/CRC Press, 2003.

[SUN2003] Sung-Bae Cho, S-B., and Won, H-H. Machine learning in DNA microarray analysis for cancer classification. Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics, Vol. 19, pp. 189-198, 2003.

[YEA2001a] Yeang, C.-H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R.M., Angelo, M., Reich, M., Lander, E.S., Mesirov, J., and Golub, T. Molecular classification of multiple tumor types. *Bioinformatics*, 17:S316-S322, 2001.

[YEA2001b] Yeang, C-H., *et al.* Molecular classification of multiple tumor types. *Bioinformatics*, 1(1):1-7, 2001.

12.6.6 Discriminatory Gene Selection

The problem of how to select the genes that best discriminate between the different disease states is well-known in the machine learning community as the problem of feature-selection. The gene-selection methods are used in order to estimate the *correlation strength* of genes that appear in important clusters with any of the samples' categories (i.e., disease-types; disease-recurrence states etc). Genes with *high ranked ordered correlation scores* will be proposed as possible indicative markers for these categories.

12.6.6.1 Available Systems and Tools

GenePattern : http://www.broad.mit.edu/cancer/software/genepattern/

12.6.6.2 References

[BAI1988] Baim, P.W. A Method for Attribute Selection in Inductive Learning Systems. *IEEE PAMI*, 10(6):888-896, 1988.

[CHE1997] Chen, Y., Dougherty, E.D. and Bittner, M.L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics*, 2:364-374,

1997.

[DUD2002] Dudoit, S., Yang, Y.-H. Callow, M.C., and Speed, T.P. Statistical methods for identifying differentially expressed genes in replicated cNDA microarray experiments. *Statistika Sinica*, 12(1), 2002.

[EFR2001a] Efron, B. Robbins, empirical Bayes, and microarrays. *Dept. of Statistics, Stanford*, Technical Report 2001-30B/219, 2001.

[EFR2001b] Efron, B., Storey, J.D., and R. Tibshirani, R. Microarrays empirical Bayes methods, and false discovery rates. *Dept. of Statistics, Stanford,* Technical Report 2001-23B/217, 2001.

[EFR2001c] Efron, B., Tibshirani, R., Storey, J.D., and V. Tusher, V. Empirical Bayes analysis of a microarray experiment. *JASA*, 96:1151-1160, 2001.

[GOL1999] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531-537, 1999.

[GOS2001] Goss Tusher, V., Tibshirani, R., and G. Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98:5116-5121, 2001.

[HAL2000] Hall, M. A. Correlation-based feature selection for discrete and numeric class machine learning. In Langley, P. (ed), *Proc Seventeenth International Conference on Machine Learning,* Stanford, CA, 359-366. Morgan Kaufmann Publishers, San Francisco, California, 2000.

[HED2001] Hedenfalk, I., *et al.*: Gene-expression profiles in hereditary breast cancer. *N Engl J Med.* 344(8):539-548, 2001.

[KOH1996] Kohavi, R., John, G. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273-324, 1996.

[LI2001] Li, L., Weinberg, C.R., Darden, T.A., Pedersen, L.G.: Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, 17(12):1131-1142, 2001.

[LOE2002] Lönnstedt, I., and Speed, T.P. Replicated microarray data. *Statistika Sinica*, 12(1), 2002.

[NEW2001] Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., and Tsui, K.W. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comp. Biol*., 8:37-52, 2001.

[POM2002] Pomeroy, S.L., *et al.*: Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436-442, 2002.

[SU2001] Su, A.I., *et al.*: Molecular Classification of Human Carcinomas by Use of Gene Expression Signatures. *Cancer Research*, 61:7388-7399, 2001.

[TIN2000] Ting Lee, M.L., Kuo, F.C., Whitmore, G.A., and Sklar, J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *PNAS*, 97:9834-9839, 2000.

12.7 Systems Biology Approaches: Molecular Pathways & Cells Modelling

An ambitious direction is to attempt to model and infer regulatory networks on a global scale, or along more specific subcomponents such as a pathway or a set of co-regulated genes. A major obstacle is that our knowledge of transcription and other critical molecular level mechanisms remains incomplete, especially as refers to in-vivo perturbations or “noise” at various stages of regulation in molecular processes which could mark the difference between changes, often epigenetic, which may significantly affect other processes, versus those

which do not. Furthermore, there are very few examples of regulatory circuits for which detailed information is available, and they all appear to be very complex.

12.7.1 Metabolic Pathways & Gene Regulatory Networks

On the theoretical side, several mathematical formalisms have been applied to model genetic networks. These range from discrete models, such as Boolean networks, as in the pioneering work of Kauffman, to continuous models based on differential equations, such as continuous recurrent neural networks or power-law formalism, probabilistic graphical models and Bayesian networks. None of these formalisms appears to capture all the dimensions of gene regulation and most of the work in this field is still very preliminary. The manual inference of pathway information as it occurs e.g. in the interpretation of gene expression data [API2005] is assisted with the use of pre-complied protein interaction databases, like those available from ingenuity, Transfac, GeneGo [NIK2005], Ariadne. A review of most of these tools can be found in [BON2004].

Understanding biology at the system level - not only gene networks, but also protein networks, signalling networks, metabolic networks, and specific systems, such as the immune system or neuronal networks - is likely to remain at the center of the bioinformatics efforts of the next few decades.

12.7.1.1 References

[API2005] Apica, G., Ignjatovicb, T., Boyerb, S. and Russell, R. Illuminating drug discovery with biological pathways. *FEBS Letters* 579:1872–1877, 2005.

[BON2004] Bonetta, L. Bioinformatics - from genes to pathways, *Nature Methods* 1(2):169, 2004.

[BRA1995] Bray. Protein molecules as computational elements in living cells. Nature, 376:307-312, 1995.

[FRI2000] Friedman, N., Linial, M., Nachman, I., Pe'er, D., Using Bayesian networks to analyze expression data. J. Comp. Biol., 7:601-620, 2000.

[MCA1999] H. H. McAdams and A. Arkin. It's a noisy business! Genetic regulation at the nanomolar scale. Trends Genet. 15:65-69, 1999.

[JEO2000] H. Jeong, B. Tomber, R. Albert, Z.N. Oltvai, and A.-L. Barabasi. The large-scale organization of metabolic networks. Nature, 407:651-654, 2000.

[YUH1998] H. Yuh, H. Bolouri, and E. H. Davidson. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. Science, 279:1896- 1902, 1998.

[HAS2000] J. Hasty, J. Pradines, M. Dolnik, and J. J. Collins. Noise-based switches and amplifiers for gene expression. Proc. Natl. Acad. Sci. USA, 97:2075-2080, 2000.

[HAR1999] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. Nature, 402, Supp.:C47-C52, 1999.

[SAV1996] M. A. Savageau. Power-law formalism: a canonical nonlinear approach to modeling and analysis. In V. Lakshmikantham, editor, World Congress of Nonlinear Analysts 92, volume 4, pages 3323-3334. Walter de Gruyter Publishers, Berlin, 1996.

[MJO1991] Mjolsness, D. H. Sharp, and J. Reinitz. A connectionist model of development. J. Theor. Biol., 152:429-453, 1991.

[NIK2005] Nikolsky, Y., Nikolskaya, N. and Bugrim, A. Biological networks and analysis of experimental data in drug discovery. *Drug Discovery Today* 10(9), 2005.

[VOI1991] O. Voit. Canonical Nonlinear Modeling. Van Nostrand and Reinhold, New York, 1991.

[KAR1999] P. D. Karp, M. Krummenacker, S. Paley, and J. Wagg. Integrated pathway-genome databases and their role in drug discovery. Trends Biotech., 17:275-281, 1999.

[KAU1969] S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. J. Theor. Biol., 22:437-467, 1969.

[KAU1990] S. A. Kauffman. Requirements for evolvability in complex systems: orderly dynamics and frozen components. Physica D, 42:135-152, 1990.

[KAU1974] S. A. Kauffman. The large scale structure and dynamics of gene control circuits: an ensemble approach. J. Theor. Biol., 44:167-190, 1974.

[YI2000] T. Yi, Y. Huang, M. I. Simon, and J. Doyle. Robust perfect adaptation in bacterial chemotaxis through integral feedback control. Proc. Natl. Acad. Sci. USA, 97:4649-4653, 2000.

[HLA1997] W. S. Hlavacek and M. S. Savageau. Completely uncoupled and perfectly coupled gene expression in repressible systems. J. Mol. Biol., 266:538-558, 1997.

[ZIE2000] Zien, R. Kuffner, R. Zimmer, and T. Lengauer. Analysis of gene expression data with pathway scores. In Proceedings of the 2000 Conference on Intelligent Systems for Molecular Biology (ISMB00), La Jolla, CA, pages 407-417. AAAI Press, Menlo Park, CA, 2000.

12.7.2 Cellular Modeling

One aim of systems biology is a better understanding of the structure and function of complex regulatory networks of biochemical reactions and gene activations. This means that in addition to studying the structure and function of single genes and enzymes, the complex relationships between biomolecules must be explored. It is noteworthy in this respect that life manifests itself through as few as a thousand different low-molecular weight substances. However, the flow of these substances is controlled by many thousands of enzymes, whose synthesis is itself controlled by a wealth of genes.

The batch of methodologies uses include: *PDE*, *Stochastic and compartmental models*, and *mixed model* approaches.

12.7.2.1 Available Systems and Tools

XPPAUT: http://www.pitt.edu/~phase/

StochSim: http://info.anat.cam.ac.uk/groups/comp-cell/StochSim.html

Discrete-event simulations: http://www-2.cs.cmu.edu/~russells/software/discrete/simulation.html

E-Cell: http://www.e-cell.org/software/ecellsystem

12.7.3 References

[DEJ2002] De Jong, H. Modeling and simulation of genetic regulatory systems: A literature review. *J. COMPUT. BIOL.* 9(1):67-103, 2002.

[WEBECE] Ecell Project. http://www.e-cell.org/

[GIL2002] Gilman, A., and Arkin, A.P. GENETIC "CODE": Representations and Dynamical Models of Genetic Components and Networks. *Annual Review of Genomics and Human Genetics*, 3:341-369, 2002.

[KIE2004] Kiehl, T., Mattheyses, R., Simmons, M., Hybrid simulation of cellular behavior. *Bioinformatics*, 20(3):316-322. 2004.

[MOR2004] Mori, H. From the sequence to cell modeling: Comprehensive functional genomics in Escherichia coli. J. *BIOCHEM. MOL. BIOL.* 37(1):83-92, 2004.

[MOR1998] Morton-Firth, C. J., and Bray, D. Predicting temporal fluctuations in an intracellular signalling pathway. *J. Theor. Biol.* 192:117-128, 1998.

[PUC2004] Puchalka, J., Kierzek, A. Bridging the gap between stochastic and deterministic regimes in the kinetic simulations of the biochemical reaction networks. *BIOPHYS. J.* 86(3):1357-1372, 2004. [*refers to 'Mixed Models' in cellular modelling*]

[TAK2003] Takahashi, K., Ishikawa, N., Sadamoto, Y., Sasamoto, H., Ohta, S. , Shiozawa, A., Miyoshi, F. , Naito, Y., Nakayamam, Y., M.Tomita. E-Cell 2: Multi-platform E-Cell simulation system. *Bioinformatics*, 19(13):1727-1729, 2003. [*refers to 'Mixed Models' in cellural modelling*]

[WEI2003] Weitzke, E., Ortoleva, P. Simulating cellular dynamics through a coupled transcription, translation, metabolic model. *COMPUT. BIOL. CHEM.* 27(4-5):469-480, 2003.

[XIA2003] Xia, X., Wise, M. DiMSim: A discrete-event simulator of metabolic networks. J. *CHEM. INF. COMP. SCI.* 43(3):1011-1019, 2003. [*refers to 'Stochastic & Compartmental Model' in cellular modelling*].

13 Tools for the visual orchestration of services

One of the main ideas within ACGT is to use workflow engines to automatically access distributed resources through machine-machine (web service) interfaces. Alongside this there may obviously still be a niche for orchestrating the use of web-based interfaces (through their "human" interfaces). From the usability point of view for non-IT proficient users the need for the implementation of visual tools that will assist them in locationg, identifying the specific characteristics of these services (through their metadata) their invocation and orchestration of paramount importance.

This Chapter provides a review of visual techniques and tools supporting invocation and orchestration of services for data access, analysis, visualisation etc, while explicitly excluding the domain of scientific workflow systems.

Given the expectation that all key information used within ACGT will be delivered through the Web, we have focussed explicitly on issues relating to Web-based resources. Thus, for example, we have not surveyed toolkits supporting end-user composition of units that are assumed to be local. We have also assumed that the existence and location of desired resources are known in advance; we have not surveyed techniques for searching for resources.

The emphasis on visual techniques arises from the need that, in ACGT, end users who are not expert programmers will be able to create, modify and share combinations of information and services. That said, there are many levels of end-user development ranging from simple tailoring of existing tools through to programming from scratch. Within this range, many scientists have already shown that they are willing and able to master scripting languages such as Perl or Python, be it at the level of fragments of 'glue code', needed to invoke and coordinate other tools, or at the level of complex scripts. With these continua in mind we have not restricted this survey to tools aimed explicitly at the most naïve of end users.

The remainder of the chapter consists of the following sub-sections:

10.1 Resource types

A description of the broad categories of resource types considered in this survey

10.2 Access and extraction

Mechanisms for accessing resources, either in their entirety or with extraction of some designated portion

10.3 Gathering and browsing

Mechanisms for bringing information together in one place and browsing it

10.4 Orchestration

Mechanisms for automating the use of multiple resources, including the use of information obtained from one resource as input to others

10.5 Republishing

Mechanisms for a user to make local information, including novel compositions of data and/or services, available in turn over the Web

10.6 General issues

Testing and validation: How a user can confirm that a tool is accessing the intended resources, and processing them in the intended way.

History and provenance: How a user's activities can be archived to enable revisiting of useful explorations

13.1 Resource types

We expect that any resource to be used within ACGT will be of one of the following types:

- **Web document**

  By this we refer to anything accessible in terms of a URI, for example requested through a Web browser. Thus these resources include HTML pages, of course, but also plain-text files, image files, XML, PowerPoint presentations etc.

- **Web application**

  A Web application is also primarily designed to be accessed through a Web browser, but typically offers the chance for the user to parameterise the request by specifying values for input fields. Such requests can also be faked by automated processes, allowing Web applications to be invoked without a user present.

- **Web service**

  A Web service is designed specifically to be accessed by automated processes, by supplying inputs and receiving results according to a declared 'service definition'.

- **Database or other legacy resource**

  For various reasons an ACGT user may need to work with resources that are not ordinarily accessible using the client/server protocols of the Web. However, we have assumed that in ACGT most such resources will be transformed into one of the other three types – for example, by having the owner of a database build a Web application providing appropriate access. Where this is not feasible, for example in cases where such access would represent an unacceptable security risk, special alternative measures (such as access through an application's API) may be required.

13.2 Access and extraction

In this section we consider techniques for accessing a resource, either in its entirety or to return just some specified portion – for example, the textual contents of a given table cell within an HTML page. These techniques apply to Web documents and Web applications.

The simplest kinds of manual access are those supported by the operations of a standard Web browser, such as Internet Explorer or Netscape. Such access includes loading a page

or file designated by a known URI, navigating through a collection of HTML pages by manually clicking on links, and driving a Web application by supplying values for input fields.

In the absence of a browser, resources can be accessed by issuing commands in one of various scripting languages that include extensions or libraries specifically for Web access. The language Perl, for example, offers Web-related commands through the library called WWW:Mechanize (also known as Mech). The current version is described at http://search.cpan.org/~petdance/WWW-Mechanize-1.18/. Such commands can be issued one by one in a command shell, or grouped into scripts. Clearly some programming proficiency is necessary to understand and to write such commands. The language WebL [KIS1998], designed explicitly for Web document processing, includes several features that simplify the fetching, processing and generation of HTML pages – but it, too, has a formal syntax that would make it inaccessible to non-professional programmers.

On the other hand, some languages have been designed with the explicit intention that they should be usable by non-professional programmers. For example, the extensions to JavaScript embodied by Chickenscratch, the scripting language for the browser plug-in Chickenfoot [BOL2005b], include commands that let users refer to portions of HTML pages using simple English-like expressions. The expression language comes from previous work referred to as LAPIS (described by Miller [MIL2003] as part of a brief but useful overview of end-user issues in Web programming). A further characteristic that simplifies the use of Chickenscratch as opposed to other script-like languages is that because Chickenfoot runs within a Web browser it can rely on sophisticated facilities provided by the browser itself, such as the handling of sessions and cookies, and security.

The Web access and extraction facilities of the agent language Visual AgenTalk in the AgentSheets environment [REP2003] are also designed for simplicity of specification by users. The result – Agent-Based Multimodel Web Interaction – is described as a move towards a Pragmatic Web, where users can easily overlay their own interpretations on values extracted from Web resources, such as a planning decision based on temperature and wind readings from a weather site. However, it is not clear that the Visual AgenTalk facilities are more than a proof-of-concept for such potential simplification.

An alternative to expressing Web access with explicit commands is for the user to demonstrate the navigation and extraction that is wanted, and have the system record what has been demonstrated for later re-play. The so-called smart bookmarks in WebViews [FRE2001] are created using a VCR-style interface that records the user's navigation and form-filling actions in using a Web application. At the time of that description of WebViews, the user could specify which portion(s) of an application's result page should be extracted, by writing expressions in the powerful and flexible XPath expression language. The authors admitted that the syntax of XPath expressions would be quite daunting for end users, so they were working on a point-and-click interface to ease that part of the task. C3W [FUJ2004] includes a point-and-click interface for specifying at least the location of elements to be extracted. While C3W does not support the full generality of the XPath match conditions used by WebViews, at least it thus allows simple use of a Web application to be recorded purely by visual actions. The IEPAD (Information Extraction based on Pattern Discovery) data extractor reported by Hsu et al [HSU2005] provides interactive techniques for the user to inspect and edit an XPath-style pattern until it extracts precisely the desired elements; these patterns are then incorporated in a record of the user's navigation in a specialised Web Navigation Description Language (WNDL). The Trainable Information Agents (TrIAs) [BAU2000] in the InfoBeans environment [BAU1999] offer what appears to be a more user-friendly interface for editing extraction patterns, in that the user can choose and refer to 'context' and 'landmark' points within the original page. Thresher [HOG2005], designed for

use with the Haystack semantic-web browser, is an attempt to specify extraction purely by demonstration of candidate matches within the HTML page itself.

Independent of such user-interface issues, the general problem of building so-called wrappers that will extract desired portions of a semi-structured resource such as an HTML page has received a great deal of attention. A 2002 survey by Laender et al [LAE2002] still appears to be the most comprehensive classification and explanation of the various approaches used, while the survey by Kuhlins and Tredwell 0 extends the coverage to commercial tools. One of the perennial problems for wrappers is that they may cease to extract the intended information if the format of the source page happens to change; Amorphic [GRE2006] is one of many attempts to build resilience to such changes into the wrappers themselves.

13.3 Gathering and browsing

As a corner of the vast research field of Personal Information Management, a small class of systems have been developed to support a user in gathering and organising a collection of information extracted from Web resources. We briefly mention some of these systems.

HunterGatherer [SCH2002] is 'an interface that lets Web users carry out three main tasks: (1) collect components from within Web pages; (2) represent those components in a collection; (3) edit those component collections.' Though these goals may not appear ambitious at first glance, the cited paper provides valuable details of issues that arise in achieving them. Internet Scrapbook [SUG2000] supports a Programming-by-Example approach to allow users to create collections of clips taken from pages that are expected to be updated frequently; the clips therefore include the extraction information necessary to re-locate the equivalent information on the updated page.

Haystack [QUA2004], mentioned above, pre-dates the Semantic Web but has been re-cast as an early example of a Semantic Web Browser – in other words, a browser for navigating a web of resources that have not necessarily been linked explicitly by the information providers, but are discovered to be related because their content has been given appropriate semantic markup. For example, when specifying a city to discover details on how to travel there, it should become straightforward to access other resources associated with that city, such as its current weather, cultural events scheduled to take place there, and so on. The mSpace project [SCH2005b] additionally considers the mechanisms by which a user could build a personal Semantic Web that includes his or her own specified associations between resources.

13.4 Orchestration

This term is sometimes used interchangeably with **choreography**, though an emerging distinction is described in the glossary at looselycoupled.com, a site dealing with web-service issues:

> ***orchestration****: Co-ordination of events in a process. Overlapping with the related concept of choreography, orchestration directs and manages the on-demand assembly of multiple component services, to create a composite application or business process. Orchestration tends to imply a single co-ordinating force, whereas choreography also applies to shared co-ordination*

> *across multiple autonomous systems. After evaluating several competing specifications, mainstream sentiment is now converging on BPEL4WS as the core standard for web services orchestration.*

We first consider the orchestrated use of Web documents and Web applications. Many of the tools listed under 'Access and Extraction' include facilities for accessing one resource on the basis of results obtained from another. The tools have differing levels of expressibility and differing levels of ease-of-use. Tools based on scripts written in a full-fledged programming language, such as Perl or the Web-browser-enabled Chickenfoot, tend to offer the greatest expressibility but may be too complex for most end users. On the other hand, while the simplified language abstractions in a tool such as AgentSheets may appear easier to understand, for anything beyond the simplest kinds of orchestration there may be problems of scaling, such as the user being unable to predict whether a setup involving multiple agents will have the combined behaviour that was intended.

Some tools offer a simple way to transfer results from one resource into inputs needed for another, such as in Lixto [BAU2003], InfoBeans [BAU1999], and Ito and Tanaka's visual environment [ITO2003]. To cope with cases in which the values cannot be transferred as-is but need some processing (for example, to convert from one format to another), other tools are based on a spreadsheet-like programming model in which the connections are expressed as formulas. One spreadsheet-like research prototype is A1 [KAN2005], which offers both formula-style and event-driven (hence agent-like) behaviour. For the time being A1 is specialised to the domain of system administration rather than Web-resource access. However, its support for arbitrary Java scripts suggests that it could easily be extended to perform Web-resource operations, albeit at the cost of requiring users to work with Java's syntax. The RecipeSheet [LUN2005] is a spreadsheet-like tool in which derivations can be written in end-user-level languages such as Rexx for its inter-cell derivations. On a RecipeSheet it is also possible to connect directly by visual wiring if a direct transfer of values is all that is needed. Adieu [SHI2005], a research prototype downloadable through IBM's alphaWorks Web site, also appears to support spreadsheet-style formulas in a highly visual interface that is intended to let non-professional programmers create their own Web applications and Web services by orchestrating others. Adieu appears easy to use for at least simple resource orchestration. A more specialised orchestration tool, again designed for 'users of spreadsheet-level ability', is WebFormulate [LEO2002], which is described as a visual continual query system for formulating temporal ad hoc analyses over networks of heterogeneous, frequently-updated data sources. The techniques of WebFormulate may well be of value in ACGT.

By contrast, the bulk of the technologies to support orchestration of Web services are designed for professional programmers. In particular, the technologies used to create web-service integrations known as mash-ups (e.g., see http://www.programmableweb.com/) fall under the umbrella term Web 2.0, a soup of programming languages and standards. The state of Web 2.0 in late 2005 is described by O'Reilly [ORE2005].

In addition there is a great deal of interest in the automated discovery and orchestration of Web services to support online business. This involves challenges at many levels; [PEL2003a] is a 2003 survey of some of the key technologies, while attempts to explain the remaining complexities and what pieces are still missing. Recent activity centres on standards such as BPEL4WS, also known as BPEL (http://xml.coverpages.org/coordination.html#bpel). Meanwhile the WfMC (Workflow Management Coalition) maintains a page at http://www.wfmc.org/standards/XPDL.htm listing tools that conform to the XPDL (XML Process Definition Language) standard.

Much of the work in this area is proprietary, but some tools listed by the WfMC, such as Nautica (http://nautica.sourceforge.jp/index-e.html) and Enhydra (http://www.enhydra.org/workflow/shark/index.html), are available under LGPL (the GNU Lesser General Public License). There is also a list of open-source workflow systems at http://joe.lindsay.net/openworkflow.html. However, we repeat that for the time being we have only been able to find tools aimed at professional programmers.

Though all but drowned out by the noise from industry, academic research into Web-service orchestration is of course also under way. The project WRABBIT, for example, is addressing how to solve problems relating to flexible construction of orchestration, and how to deal with errors in inter-service conversations. And researchers in bioinformatics have already built up some experience on the potential uses of Web-service orchestration in their domain.

13.5 Republishing

Here we consider how users can deliver their local resources, which may include locally built assemblies of remote resources, as Web documents, applications or services. As an exercise in collaborative science, ACGT is likely to include many scenarios where researchers must contribute data and services to the corpus of available resources, rather than just acting as consumers.

An overview of this area is provided by Rode et al's 2004 [ROD2004] survey of commercially available tools for end user development for the Web. The authors identify eleven 'problem areas' for such development, as follows (from Table 2 in [ROD2004]):

1. Getting Started – support provided during the initial phases of development
2. Workflow – General process for creating a web application
3. Level of Abstraction – Abstractions provided by the tool, e.g., components, wizards
4. Layout – Defining the look of the web application
5. Database – Creating and accessing a database
6. Application Logic – Defining the behaviour of the web application
7. Testing and Debugging – Helping to identify and solve errors
8. Learning and Scaling – Support learning and growing beyond the 'ceiling' of the tool
9. Security – Level of security of produced web applications
10. Collaboration – Supporting multiple developers
11. Deployment – Assisting in moving a ready application into production mode.

The sheer range of issues to be addressed may explain why there do not seem to be many non-commercial tools offering more than basic, isolated functionality. We mention a few such projects. WebSheets [WOL2002] was a proposal for a 'programming in the interface' approach to creating simple HTML interfaces to simple databases. The research prototype FAR [BUR2001] was more ambitious, based on an end-user language combining ideas from web-page layout, spreadsheets, and rules to support definition of e-services by non-professional programmers. IBM's Adieu [SHI2005], mentioned above, provides simple mechanisms for exporting newly created compositions as Web applications or services. The most comprehensive approach so far, though evidently still in its early stages, seems to be the project Click, developed by the group responsible for the above survey. One of the key

contributions of this project is its emphasis on a 'gentle slope of complexity' for development, encompassing the following levels (from Figure 5 of [ROD2005]):

1. Customising template web applications
2. Using Wizards to create related sets of components
3. Designing via WYSIWYG, direct manipulation, parameter forms
4. Editing layout code (similar to HTML, ASP.NET, JSF)
5. Editing high-level behaviour code
6. Modifying and extending the underlying component framework
7. Editing PHP code

We suggest that this list may serve as a useful reference for deciding what level of user development is appropriate in any ACGT scenario involving end user development. However, since the current incarnation of Click is focussed on construction of 'basic data collection and management' applications, where the resources under its control are just simple local databases, Click is unlikely to be directly applicable in ACGT unless and until it is extended to manage Web resources too.

13.6 General issues

Finally we consider some issues that cut across the structure of this survey.

13.6.1 Testing and validation

While ACGT has acknowledged the need for rigorous validation of new software components developed within the project, it must also recognise that assemblies of software by end users introduce their own opportunities for errors, failures, and security risks. Panko's often-quoted and conscientiously updated survey of errors in spreadsheets [PAN2005] stands as a sobering reminder of the dangers of providing programming facilities to users who, not being trained in software engineering, fail to appreciate the likelihood and the potential consequences of code errors. This has led to a call for a new practice of End User Software Engineering, with groups such as Burnett et al [BUR2004] suggesting ways to encourage and guide users into checking the correctness of their work.

13.6.2 History and provenance

Schraefel and her colleagues at the University of Southampton have observed that ready and flexible access to bioinformatics resources on the Web has not been accompanied by the development of an electronic equivalent of a lab book in which scientists can conveniently record, for future reference, what they have done and how it turned out. They have therefore embarked on a project to record 'context histories' [SCH2005a] that would support this goal, the latest incarnation being myTea [GIB2006]. The group appear to be still in the early stages of defining how myTea should work alongside the history and provenance services available in Grid-based science toolkits, but we believe that the initiative is highly relevant to ACGT's goals in supporting exploratory science, and should be followed closely by the project.

13.6.3 Related EU Projects

13.6.3.1 EUD-NET: End-User DevelopmENT

Empowering people to flexibly employ advanced information and communication technology Thematic network contract IST-2001-37470 July 2002 – Aug 2003 (http://giove.cnuce.cnr.it/eud-net.htm).

'The main purpose of the network of excellence was to help the European Commission to prepare a research agenda in the end-user development field for the next framework and to increase contacts among highly-qualified research centres, both academic and industrial, in order to ease exchange of information and results and speed-up the production of innovative ideas and approaches in the field considered.'

The deliverable 'Research Agenda and Roadmap', available at http://giove.cnuce.cnr.it/EUD-NET/pdf/Research%20Agenda%20&%20Roadmap.pdf, provides useful context for ACGT's efforts to support non-programming users.

13.6.3.2 MIND

Resource selection and data fusion for Multimedia INternational Digital libraries (http://www.mind-project.org/) Cost-sharing contract IST-2000-26061 Jan 2001 – June 2003

'The key objective of the MIND project was to address the problems faced by users in terms of their ability to access and exploit the increasing number of digital libraries available internationally through networks, like the Internet and the World Wide Web (WWW). More specifically MIND aimed at designing models and building set of tools and associated test-beds to improve the effectiveness of resource selection, multimedia information access, retrieval and fusion of the retrieval data.'

13.6.3.3 PSI3: Personalised Services for Integrated Internet Information

(http://www.psi3project.org/) Cost-sharing contract IST-1999-11056 Feb 2000 – July 2002

'PSI3 addressed the problem of the exponential growth of information services on the Internet. PSI3 investigated innovative techniques to personalise and integrate existing Internet information. It analysed two different solutions to gather information, a set of improvements over existing information analysis, indexing and processing, i.e. new paradigms on content-based multimedia data retrieval and a set of personalisation features. The results of this project were firstly, a set of integrated tools accesible to application developers by an SDK, and secondly, three end-user applications which were evaluated and validated the PSI3 technical achievements.'

It is not clear, from the web site at least, what results were obtained by this project.

13.6.3.4 MINING MART: Enabling End-User Data warehouse Mining

(http://mmart.cs.uni-dortmund.de/) Cost-sharing contract IST-1999-11993 Jan 2000 – Sept 2002

'The project aimed at new techniques that give decision-makers direct access to information stored in databases, data warehouses, and knowledge bases. The main goal was the integration of data and knowledge management. Discovery techniques produce knowledge

from very large sets of distributed data. They exploited domain knowledge in order to deliver more concise and relevant insights. The main obstacle to achieve this goal was the problem of finding the proper representation for a discovery task.'

13.7 References

[BAU1999] Bauer M, Dengler D. InfoBeans - Configuration of Personalized Information Services. In: The Intl Conf on Intelligent User Interfaces (IUI '99), Redondo Beach, CA, USA, 1999, 153-156.

[BAU2000] Bauer M, Dengler D, Paul G. Programming by Demonstration for Information Agents. In H. Lieberman (Editor), Your Wish is My Command: Giving Users the Power to Instruct their Software. Morgan Kaufmann Publishers, 2000, 87-114.

[BAU2003] Baumgartner R, Gottlob G, Herzog M. Visual programming of web data aggregation applications. In Proceedings of the Eighteenth international joint conference on articial intelligence. Workshop on Information Integration on the Web (IIWeb-03), Acapulco, Mexico, 2003, 137-142.

[BLA2005a] Blanchet W, Elio R, Stroulia E. Conversation Errors in Web Service Coordination: Run-time Detection and Repair. In: ACM Intl Conf on Web Intelligence (WI 2005), 19-22 Sept 2005, Compiegne, France, 442-449.

[BLA2005b] Blanchet W, Stroulia E, Elio R. Supporting Adaptive Web-Service Orchestration with an Agent Conversation Framework. In: IEEE Intl Conf on Web Services (ICWS'05), 541-549, 2005.

[BOL2005a] Bolin M. End-user Programming for the Web. MEng thesis, Massachusetts Institute of Technology, June 2005. Available at http://groups.csail.mit.edu/uid/projects/chickenfoot/mbolin-thesis.pdf

[BOL2005b] Bolin M, Webber M, Rha P, Wilson T, Miller R C. Automation and customization of rendered web pages. In: The 18th Annual ACM Symposium on User interface Software and Technology (UIST '05), Seattle, WA, USA, Oct 23-26, 2005, 163-172.

[BUR2001] Burnett M, Chekka S, Pandey R. FAR: An End-User Language to Support Cottage E-Services. In: 1st IEEE Symp. on Human-Centric Computing Languages and Environments, 2001, 195-202.

[BUR2004] Burnett M, Cook C, Rothermel G. End-user software engineering. Communications of the ACM (Sept 2004) 47(9):53-58.

[FRE2001] Freire J, Kumar B, Lieuwen D. WebViews: accessing personalized web content and services. In: The 10th international Conference on World Wide Web (WWW '01), Hong Kong, May 1-5, 2001, 576-586.

[FUJ2004] Fujima J, Lunzer A, Hornbæk K, Tanaka Y. Clip, Connect, Clone: Combining Application Elements to Build Custom Interfaces for Information Access. In ACM UIST 2004, Santa Fe, NM, 175--184.

[GIB2006] Gibson A, Stevens R, Cooke R, Brostoff S, schraefel m c. myTea: Connecting the Web to Digital Science on the Desktop. Submitted to World Wide Web Conference 2006, Edinburgh. Available at http://eprints.ecs.soton.ac.uk/11549/01/www2006myTeaFinalSubmission.pdf

[GRE2006] Gregg D G, Walczak S. Adaptive web information extraction. Communications of the ACM (May 2006) 49(5):78-84.

[HOG2005] Hogue A, Karger D. Thresher: automating the unwrapping of semantic content from the World Wide Web. In: The 14th international Conference on World Wide Web (WWW '05), Chiba, Japan, May 10-14, 2005, 86-95.

[HSU2005] Hsu C-N, Chang C-H, Hsieh C-H, Lu J-J, Chang C-C. Reconfigurable Web wrapper agents for biological information integration. Journal of the American Society for Information Science and Technology, Mar 2005, 56(5):505-517.

[ITO2003] Ito K, Tanaka Y. A visual environment for dynamic web application composition. In: The Fourteenth ACM Conference on Hypertext and Hypermedia (HYPERTEXT '03), Nottingham, UK, August 26-30, 184-193, 2003.

[KAN2005] Kandogan E, Haber E, Barrett R, Cypher A, Maglio P, Zhao H. A1: end-user programming for web-based system administration. In: The 18th Annual ACM Symposium on User interface Software and Technology (UIST '05), Seattle, WA, USA, October 23-26, 2005, 211-220.

[KIS1998] Kistler T, Marais H. WebL - a programming language for the Web. In: The Seventh international Conference on World Wide Web 7 (WWW '98), Brisbane, Australia, 1998, 259-270.

[KUH2003] Kuhlins, S. and Tredwell, R. Toolkits for Generating Wrappers - A Survey of Software Toolkits for Automated Data Extraction from Web Sites. Lecture Notes in Computer Science (LNCS) 2591:184-198. 2003.

[LAE2002] Laender A, Ribeiro-Neto B, Silva A, Teixeira J. A Brief Survey of Web Data Extraction Tools. SIGMOD Record (June 2002), 31(2):84-93.

[LEO2002] Leopold J, Heimovics M, Palmer T. WebFormulate: a web-based visual continual query system. In: The 11th international Conference on World Wide Web (WWW '02), Honolulu, Hawaii, USA, May 7-11, 2002, 221-231.

[LOR2004] Lord P, Bechhofer S, Wilkinson M D, Schiltz G, Gessler D, Hull D, Goble C, Stein L. Applying semantic web services to bioinformatics: Experiences gained, lessons learnt. In International Semantic Web Conference, 2004, 350-364.

[LUN2005] Lunzer A, Hornbæk K. An Enhanced Spreadsheet Supporting Calculation-Structure Variants, and its Application to Web-Based Processing. In K.-P. Jantke, A. Lunzer, N. Spyratos and Y. Tanaka (eds.) Proceedings of the Dagstuhl Workshop on Federation over the Web, Dagstuhl Castle, Germany, May 2005, (Lecture Notes in Artificial Intelligence, Vol. 3847/2006), 143-158.

[MIL2003] Miller R C. End User Programming for Web Users. In Workshop on End User Development at the Conference on Human Factors in Computer Systems (CHI 2003), April 2003. Available at http://giove.cnuce.cnr.it/chi/doc/eup.pdf

[ORE2005] O'Reilly T. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. 30 Sept 2005. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[PAN2005] Panko R R. What We Know About Spreadsheet Errors. Jan 2005. Available at http://panko.cba.hawaii.edu/ssr/whatknow.htm

[PEL2003a] Peltz C. Web Services Orchestration: A review of emerging technologies, tools, and standards. Jan 2003. Available at http://devresource.hp.com/drc/technical_white_papers/WSOrch/WSOrchestration.pdf

[PEL2003b] Peltz C. Web Service Orchestration and Choreography: A look at WSCI and BPEL4WS. July 2003. Available at http://devresource.hp.com/drc/technical_articles/wsOrchestration.pdf

[PET2003] Petrie C, Bussler C. Service Agents and Virtual Enterprises: A Survey. Internet Computing, July/August 2003, 7(4). Available at http://snrc.stanford.edu/~petrie/fx-agents/xserv/icpaper/

[QUA2004] Quan D, Karger D R. How to Make a Semantic Web Browser. In: The Thirteenth International World Wide Web Conference (WWW 2004), New York City, May 17-22, 2004.

[REP2003] Repenning A, Sullivan J. The Pragmatic Web: Agent-Based Multimodal Web Interaction with no Browser in Sight. In: IFIP Conf. Human-Computer Interaction (INTERACT '03), 2003, 212-219.

[ROD2005] Rode J, Bhardwaj Y, Pérez-Quiñones M A, Rosson M B, Howarth J. As Easy as "Click": End-User Web Engineering, Lecture Notes in Computer Science, July

2005, 3579:478-488

[ROD2004] Rode J, Howarth J, Pérez-Quiñones M A, Rosson M B. An End-User Development Perspective on State-of-the-Art Web Development Tools. Technical Report TR-05-03, Computer Science, Virginia Tech. 2004

[SCH2005a] Schraefel m c, Brostoff S, Cooke R, Stevens R, Gibson A. Transparent interaction; dynamic generation: context histories for shared science. In: Proceedings of workshop ECHISE 2005 - 1st International Workshop on Exploiting Context Histories in Smart Environments held in Conjunction with The Third Internationl Conference on Pervasive Computing 2005, Munich, Germany, May 11 2005.

[SCH2005b] Schraefel m c, Smith D A, Owens A, Russell A, Harris C, Wilson M. The evolving mSpace platform: leveraging the semantic web on the trail of the memex. In: The Sixteenth ACM Conference on Hypertext and Hypermedia (HYPERTEXT '05), Salzburg, Austria, Sept 6-9, 2005, 174-183.

[SCH2002] Schraefel m c, Zhu Y, Modjeska D, Wigdor D, Zhao S. Hunter Gatherer: Interaction Support for the Creation and Management of Within-Web-Page Collections. In: The 11th International World Wide Web Conference (WWW2002), Honolulu, Hawaii, USA, 2002, 172--181.

[SHI2005] Shinomi H, Adams S S. Say goodbye to complexity when developing Web services (ADIEU: The End User Computing Tool for Web applications and Web services). 6 Oct 2005. Available at http://www-128.ibm.com/developerworks/web/library/ws-adieu/index.html

[SUG2000] Sugiura A. Web Browsing by Example. In H. Lieberman (Editor), Your Wish is My Command: Giving Users the Power to Instruct their Software. Morgan Kaufmann Publishers, 2000, 61-85.

[WOL2002] Wolber D, Su Y, Chiang Y T. 2002. Designing dynamic web pages and persistence in the WYSIWYG interface. In: The 7th international Conference on intelligent User interfaces (IUI '02) San Francisco, CA, USA, Jan 13-16, 2002, 228-229.

[ZHA2005] Zhao J, Lord P, Alper P, Wroe C, Goble C. The implications of semantic web technologies for support of the e-Science process. In Proc UK e-Science All Hands Meeting 2005. EPSRC, 2005.

14 Approaches to the Integration of Heterogeneous Databases

14.1 Introduction

According to the workplan, the main challenge in ACGT is the semantic integration of heterogeneous databases managing multilevel (phenotypical and genotypical) data. As a result in this section we review state-of-the-art in the domain of integration strategies between genomic, proteomic, clinical and population databases.

The process of heterogeneous database integration may be defined as “the creation of a single, uniform query interface to data that are collected and stored in multiple, heterogeneous databases.” [SUJ2001].

In biomedicine and other domains, the problems of heterogeneous database integration are being addressed in research and applications environments. This Chapter reviews the nature of heterogeneous biomedical database integration and the general methodologies that researchers have pursued to overcome the problem.

14.2 The heterogeneous database integration problem

In recent years, there has been an enormous growth in the number of publicly accessible databases on the Internet. All indications suggest that this growth will continue in the years to come. Access to these data presents several complications and problems.

The first complication is *distribution*. Many queries will not be answered by providing data from a single database. Useful relations and data may be broken into fragments that are distributed among distinct databases. Database researchers distinguish among two types of fragmentation; *horizontal* and *vertical fragmentation*. Distributed databases can exhibit mixtures of these types of fragmentation. Later, we will see more information about these types of fragmentation and will discuss more about the problem that this kind of division raises.

A second complication in database integration is *heterogeneity*. This heterogeneity may be **notational** or **conceptual**. *Notational heterogeneity* concerns the access language and protocols. One source might use a DBMS using a concrete query language while another source uses the same DBMS but with a different query language. A third example might use, too, a complete different DBMS and query language. This sort of heterogeneity can usually be handled through commercial products.

However, even if we agree that all the databases in a distributed system use a standard hardware and software platform, language and protocol, there can still be a *conceptual heterogeneity* as differences in their relational schemas and vocabulary. Distinct databases may use different words to refer to the same concept, and/or they may use the same word to refer to different concepts. Reassembling the distributed fragments of a database in the face of heterogeneity might prove difficult.

14.2.1 Types of heterogeneity

In trying to integrate diverse computer systems, several problems might appear, which should be solved along the integration process. Some of them are due to heterogeneity wrt 'physical' issues (hardware components or architectures, operating systems, database management systems…) and others to 'logical' issues, such as the data or representational model used to store the data. The different barriers that have to be addressed due to heterogeneity of the systems and information have been classified into 3 groups: *HW/SW Heterogeneity*, *Data Model Heterogeneity* and *Representational Heterogeneity*. The differences between them are explained in depth below.

14.2.1.1 HW / SW Heterogeneity

Integration issues can be discussed according to various criteria. From a low-level to a high-level point of view, these problems range from different computer architectures to different database management systems, including the several operating systems.

In the lowest-level (physical), a very well-known cause of conflict is about the internal data representation format used by the specific machines. Some of them use *big-endian* representation like mainframes, Mac and SUN, whereas PCs (Intel) or UNIX Servers are *little-endians.* This difference depends on whether the first location in the memory is the least or most significant byte. In the figure below illustrates better the difference.

With regard to software heterogeneity many differences can be found, according to Operating Systems or Database Management Systems (DBMS). The process are not the same or do not perform in the same way when they are running over *Windows*, *Linux* or *UNIX* platform. Each Operating System has its own procedures and functions to do tasks. The same happens with DBMS. Depending on the concrete DBMS used, data will be represented, stored, accessed and retrieved in different ways.

14.2.1.2 Data model heterogeneity

This type of heterogeneity has been outlined and come derived from some aspects mentioned on previous section. Before integrating any system, it is needed to take into account the data and format of these data. Very often it is assumed that data are already stored on databases, but this may be a great mistake because some systems might use flat files, even unstructured. Maintenance and access to this information will be completely different from "traditional" databases.

Even if DBMS are used in the systems to be integrated, important heterogeneities are useally found due to different data models employed, e.g. relational, object-oriented or hierarchical model. Data are stored and accessed in different ways depending on the concrete model. And even more, although the systems might employ the same data model, depending on the DBMS used, the syntax to handle data might be quite different.

14.2.1.3 Representational heterogeneity

The largest barrier to heterogeneous database integration is the variety with which similar data are represented in different databases, i.e., representational heterogeneity. It is appropriate to consider several types of representational heterogeneity that schema integration techniques must resolve. The most general type of heterogeneity is related to the data models themselves. Aggregating data from relational, hierarchical, object-oriented, and flat file databases into a single representation is the first step in schema integration. However, even if all database systems were to use the relational model, significant

representational heterogeneity would remain. Specifically, there exist various kinds of differences: structural, naming, semantic and content.

14.2.1.3.1 Structural differences

Structural differences may consist of alternative table decompositions (horizontal and vertical), differences in data versus metadata representation and differences in structured versus free-text encodings.

Alternative table decomposition

Alternative horizontal table decompositions entail different degrees of normalization that result in the same information being distributed across a varying number of tables. Alternative vertical decompositions entail different distributions of rows among one or more tables. Rows may be partitioned in certain databases across multiple tables to improve retrieval performance when most (local) queries access only a subset of all the rows.

Data vs. metadata representation

Since the relational model provides no constructs for representing type hierarchies directly, such hierarchies may be encoded in a variety of ways in relational databases. These encodings entail differences in the use of data and metadata. For example, encoding two types of serum electrolyte results, serum sodium and serum potassium, may be represented in at least three ways, involving the use of table names, field names, or field values.

Structure vs. free-text encodings

Using structured versus free-text encoding is a source of heterogeneity and entails differences in the distribution of primitive data elements across multiple fields versus concatenated in one field. Common examples include the separation or concatenation of laboratory result values and units (i.e., [“145 mg”] versus [“145,” “mg”] and the separation or concatenation of address components (i.e., [“125 Elm St., Denver, CO 80220”] versus [“125 Elm St.,” “Denver,” “CO,” “80220”]).

14.2.1.3.2 Naming differences

Naming differences are characterized by distinct lexical terms denoting the same semantic objects across database schemas. Naming differences may be shown as metadata or data differences. Metadata differences are among the simplest forms of database heterogeneity and comprise variations in the names of tables and fields, such as “Doctor” versus “Physician” or “MRN” versus “Patient ID.” The difficulty with metadata naming differences is discriminating differences that are solely syntactic from differences that represent variations in semantics. For example, “MRN” versus “Patient ID” is a semantic rather than naming difference if “Patient ID” denotes the social security number of the patient rather than the medical record number (semantic differences are addressed in the following section). Detecting these (sometimes) subtle semantic distinctions is among the most time-consuming aspects of database schema integration.

Similarly, data-naming differences are characterized by disparities among the symbols used to denote synonymous instances in heterogeneous databases. Examples include variations in naming diseases (“MI” versus “myocardial infarction”) and tests (“Na” versus “Serum Na” versus “Serum sodium”). In the biomedical domain, where nomenclature is complex (and sometimes ad hoc and often overlapping), this “vocabulary problem” is a significant issue for

any system that seeks to aggregate or compare data collected at distinct sites. A subfield of medical informatics is devoted to these terminology issues. Many vocabulary problems in biomedicine, however, go beyond syntactic (naming) differences to the more difficult issue of semantic differences.

14.2.1.3.3 Semantic differences

Semantic differences occur when the meanings of table names, field names and data values across local databases are similar but not precisely equivalent. This problem is particularly complex when the labels of tables, fields and data values are identical across databases, but their meanings are, in fact, different. To create a uniform query model that has well-defined semantics and produces accurate query results, these semantic differences must be recognized.

Semantic differences may occur when there is no one-to-one correspondence among the concepts denoted by values within local databases. Additionally, some cases may arise in which semantic differences exist among value sets that cannot be resolved by mapping. Specifically, this occurs when there is an overlapping (many-to-many) correspondence among value sets.

14.2.1.3.4 Content differences

Content differences occur when the data represented in a local database are not directly represented in another database. These data may be implicit, derivable or simply missing. Implicit data are usually constant and therefore assumed, within the environment of a local database, but cannot be assumed in the context of a global database. For example, an integrated database that provides a registry of licensed physicians must explicitly represent the specialty and board certification of each physician represented. However, if the underlying local databases are the membership registries of specialty societies, the specialties and board certifications of stored individuals will not be represented because they are implicit. In these cases, data-transformation or query-transformation processes will need to introduce values for the specialty and board certification depending on the source database.

A classic example of derivable data is the representation of zip code versus state or date-of-birth versus age. Clearly, each may be derived from the other (with some loss of information depending on the “direction” of the computation). These types of arbitrary transformations, however, illustrate the need for general-purpose functions within data-translation or query-translation software modules because declarative mappings are often not powerful enough to resolve such differences.

The problem of missing data occurs when the global schema contains an information type that is simply not available in one or more local databases. Further research is required to accurately represent the semantics of missing information in the query models of heterogeneous databases.

14.2.2 Requirements for heterogeneous database integration

It is useful to enumerate a set of requirements and assumptions to provide a context to the challenges of heterogeneous database integration. Those requirements are the following:

1. Database heterogeneity. Despite the efforts and advances in the area of standards, a single model for biomedical databases will not emerge, at least in the short term.

2. Users and applications must be able to issue complex declarative multidatabase queries. Heterogeneous database systems must provide powerful and general query capabilities to retrieve all the information pertaining to a single object or all the objects that meet a set of search criteria. The capabilities should not be rigidly linked to any particular application or information need.

3. Users and applications should not be required to know the existence, physical location, access mechanism, or schema of the underlying local databases.

4. The contents of the local databases may be autonomously and locally maintained.

5. The schemas of local databases change quickly (on average, two or three times per year). The databases are designed and maintained to meet local needs and changes are made independently of the integrated database structure.

6. Updates to local databases occur frequently, so the newest data should be accessible, too.

14.2.3 Database integration types

The process of heterogeneous database integration may be defined as “the creation of a single, uniform query interface to data that are collected and stored in multiple, heterogeneous databases.” Several varieties of heterogeneous database integration are useful in biomedicine. We can consider three kinds of database integration types:

- ***Vertical integration.*** The aggregation of semantically similar data from multiple heterogeneous sources. For example, a “virtual repository” that provides centralized access to different data that are collected and stored in databases across several places.

- ***Horizontal integration.*** The composition of semantically complementary data from multiple heterogeneous sources. For example, a system that supports complex queries across genomic, proteomic, and clinical information sources for molecular biologists or a physician workstation that provides a single interface to data stored in multiple ancillary systems.

- ***Integration for application portability.*** The standardization of access to semantically similar information at disparate sources. For example, a universal database interface for decision-support applications that allows them to be shared across institutions with no modifications to their implementations.

14.3 Approaches to heterogeneous database integration

In recent years the amount of information produced and stored in databases has increased greatly. For example the Human Genome project has led to immense amounts of data that have been collected in different databases. To manage and handle all this information, a new area in computer science has emerged: database integration.

Database integration is the area of computer science related with information exchange and gathering, usually from heterogeneous and disparate sources. It faces problems such as bringing together data with different pattern, or allowing users to access to information located in different places in a uniform manner. There are two basic approaches to database integration: centralized vs. federated.

The centralized approach relies on a central repository where all data is to be stored, called "Data Warehouse" [KIM1996]. Users will finally access data stored in such integrated database. The Data Warehouse has its own data model, which is independent from the original databases. This allows fast response to user queries, since all data are collected locally. However, it also has several drawbacks, such as the possibility of inconsistencies in the data (changes in the original sources may take time to reach the central repository), the elevated cost of maintaining the repository and the need of additional space, since the Data Warehouse is a new database. Figure 16 shows a representation of integration in a Data Warehouse.

Figure 19: A graphical representation of a Data Warehouse

The federated approach does not rely on a central repository, but leaves the data in the original sources. This is actually being more used nowadays, since it solves the problems of the centralized approach. The federated approach was first introduced by [SHE1990], and called Federated Database System (FDBS). In a FDBS, databases in the system are autonomous, and their local operations don't depend on the FDBS. Nevertheless, this first approach had some drawbacks itself, like the difficulty for updating data sources (include or remove data).

Another distributed approach, known as "mediation", was introduced by [WIE1992]. The mediator is a middleware layer between the user and the data sources. Mediation is not based in a Data Management System.

In this case we have a virtual model composed of all the databases we want to include. User queries are translated to queries on those sources, and results are merged together before presenting them to the user. This way the user *sees* a central repository containing all the data (all translations and integrations are performed transparently). This eliminates the drawbacks exposed before, however performance may be lower in this case since data must travel from the sources for each query.

Systems that rely on accessing federated data sources are called "query translation systems" as well. Figure 19 shows the general architecture of a query translation system.

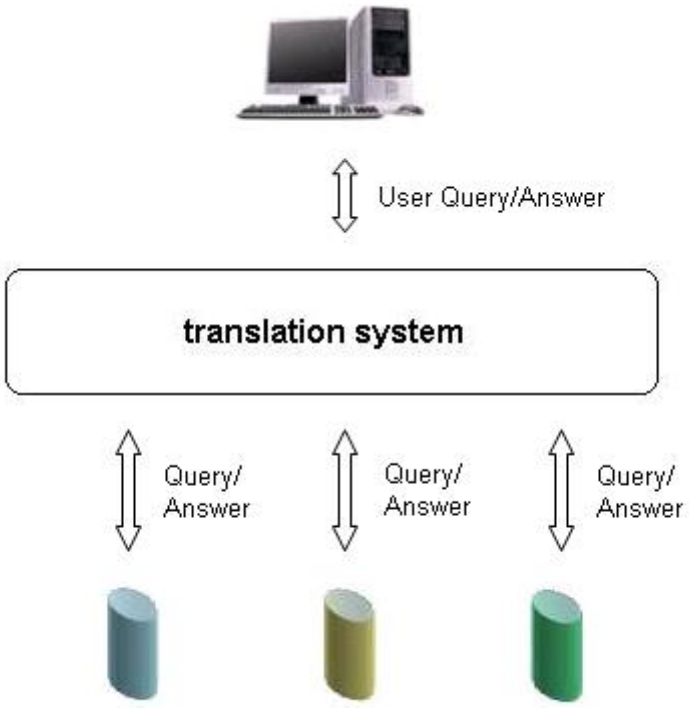


Figure 20: An example of federated database integration approach

Query translation systems can be classified in four categories: 1) pure mediation, 2) single conceptual schema approaches, 3) multiple conceptual schema approaches and 4) hybrid approaches [PER2004].

1) ***Pure Mediation Systems***

   In pure mediation systems there is no alternative data model presented to the user. Instead of that, a mediator is used to resolve user queries. A mediator is a software layer, close to the concept of middleware.

   Main drawbacks: Pure mediation systems are usually not very much intuitive.

2) ***Global Conceptual Schema Systems***

   Global Conceptual Schema Systems are based on a single ontology that models the domain of interest. Database objects are linked to objects belonging to the global ontology.

   Main Drawbacks: Addition or removal of databases may require the modification of the global ontology.

3) ***Multiple Conceptual Schema Systems***

   In the Multiple Conceptual Schema Systems each source is described using a different ontology.

   Main Drawback: It is not possible to ensure that semantically equivalent entities share names. It is required to create mappings among semantically similar objects belonging to different ontologies.

4) ***Hybrid approach***

In such Hybrid Approaches each source is described using a different ontology. Each one of these ontologies is built using objects from an ontology approved by domain experts. This led to that semantically equivalent objects belonging to different ontologies share names.

Main Drawback: Validated domain ontology is required.

14.4 Semantic Mediation

14.4.1 Introduction

A mediator is a software module that exploits encoded knowledge about certain sets or subsets of data to create information for a higher layer of applications [WIE1992]. Mediators provide:

- Transformation of databases.
- Methods to access and merge data from multiple databases.
- Abstraction and generalization of underlying data.

Mediation is close to the middleware concept. A mediator can act as a source of information for another mediator, since it provides services to access data to higher layers. Usually, a wrapper is needed for every data source, and the integration of such data sources is done under demand.

Semantic mediation aims to solve the problem of discovering data in sources of information that cannot be accessed easily. A semantic mediation system should include services for formulating semantic queries, and should give transparent access to heterogeneous sources of data.

Since ways to store data are manifold, heterogeneous database integration is a key concept in semantic mediation. The integration and access to heterogeneous sources of information can be approached in several ways, being the ontology-based approach one of the most effective, understandable and reliable.

Ontologies provide the semantics needed to save the gap between heterogeneous data sources and a formal language for information retrieval. In a semantic mediation system, the user (either if this user is a human or not) shouldn't take care about the format of the information source, but about the terms contained in the ontology for building the query in a proper way. The system would provide then a virtual view of the data, based on how ontologies describe the domain or domains implied.

14.4.2 Ontologies in information systems

There are numerous definitions for the term "Ontology". One of the most cited is that given by Gruber: "***An ontology is an explicit specification of a conceptualization***" [GRU1993]. We can also describe an ontology as what it provides: a conceptual framework for a structured representation of the meaning, through a common vocabulary, of a given domain (e.g. medical ontologies describe certain medical domain), specifying concepts, relationships between such concepts and axioms in a formal manner.

An ontology can be seen as a set of classes and the relations between them. It can be complemented with restrictions and instances of the elements belonging to the different

classes described. The following Figure shows a view of an ontology subset in the tool ONTOFUSION.

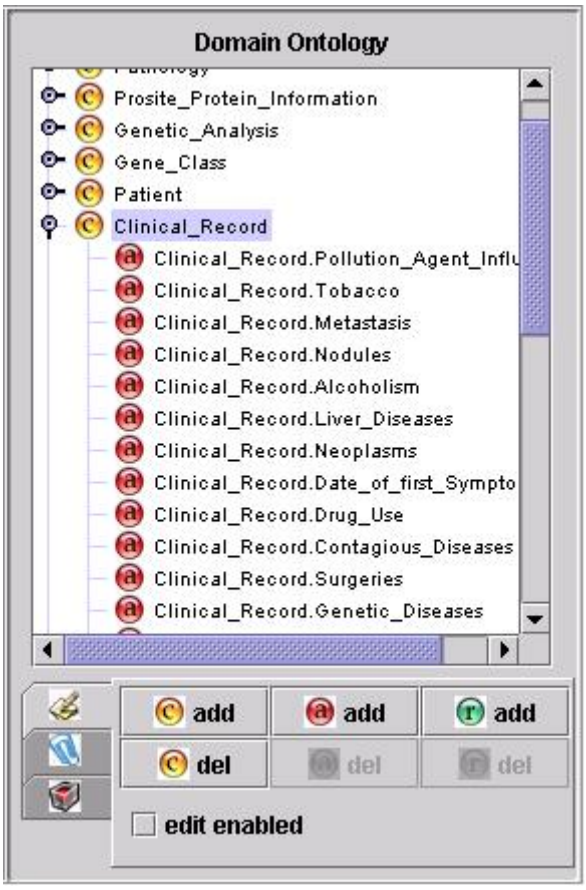


Figure 21: Example of a domain ontology in Ontofusion

There have been some approaches for defining formal ontologies, both in biological and medical domains, such as the Gene Ontology (GO), that aims to provide a controlled vocabulary to describe gene and gene product attributes in any organism, and the Unified Medical Language System (UMLS), which is a compendium of medical terms.

The real benefit from using ontologies in software development comes from the capacity of such ontologies to make data “understandable” for software entities. By having and explicit and formal definition of a given domain, our applications are able to categorize and manage data given its semantic meaning, something that was unavailable previously. On the other hand, ontologies are simple enough for humans to work with them, allowing experts to easily translate their knowledge about a given area into computer understandable knowledge.

However, we must be aware that the use of ontologies in information systems may have its drawbacks, like for example the increase in complexity of research projects and the lack of well-established standards for ontology construction and edition.

14.4.3 Ontologies applied to database integration

Database integration requires bridging the syntactic and semantic gaps existing across data sources. Given the suitability of ontologies to provide a semantic layer to applications, database integration is moving towards an ontology-based approach. This approach seems to be very promising, although there are still several issues that must be addressed.

In the biomedical domain, it has been demonstrated that ontologies can aid BI-MI integration, since they are mainly used to facilitate knowledge distribution, sharing and reuse. [PER2004]

In order to apply ontologies to database integration, several systems use ontology-based views to facilitate the mapping from objects of specific databases to shared vocabularies. There are some less common approaches, such as the use of ontologies for automatic mediator generation. Some of these systems are reviewed later in this document.

Database integration is evolving towards ontology-based approaches, where ontologies are used to support mapping between equivalent concepts for integration and query formulation, given that ontologies provide a common and shared vocabulary that can be used to facilitate the communication and information transportation between users, systems and databases.

14.4.4 Projects and International Initiatives

14.4.4.1 DataFoundry

DataFoundry [CRI1998a], [CRI2001], [CRI1998b] project aimed to improve scientists' access to distributed and heterogeneous data. The approach in this project was using ontologies for automatic mediator generation, thus reducing the efforts when including new data sources, or existing sources changed their structure.

The ontology designed for this project stored metadata about mediator generation, allowing automatic mediator creation. It included knowledge to identify and resolve both syntactic and semantic conflicts between data contained in the sources, allowing the unification of concepts contained in such data.

The ontology was composed of four basic concepts, necessary for the mediator generation:

- Abstractions: abstractions of domain specific concepts.
- Databases: database descriptions.
- Mappings: mappings between a database and an abstraction.
- Transformations: functions to resolve representation conflicts.

Figure 19 shows the architecture of DataFoundry System.

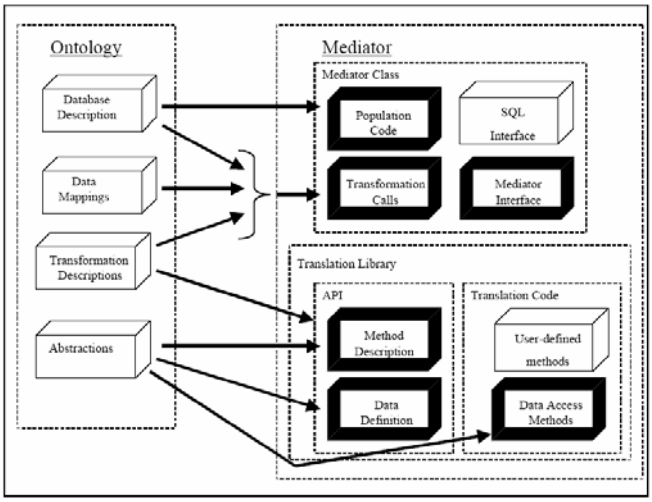


Figure 22: DataFoundry architecture

The project status is now inactive.

14.4.4.2 LinkFactory

LinkFactory [VER2003] is an Ontology Management System (OMS) that offers users a GUI for creating and managing ontologies. It was for example used in the creation of LinkBase, an ontology covering the biomedical domain. The tool offers a multiple windows environment and a series of functions to allow users easily editing and managing ontologies.

LinkFactory includes and extension tool called MaDBoKS, which allows mapping external databases to ontologies. This way, any relational schemata can be mapped with an ontology, and use this information for integrating distributed heterogeneous data sources. Figure 20 shows LinkFactory GUI.

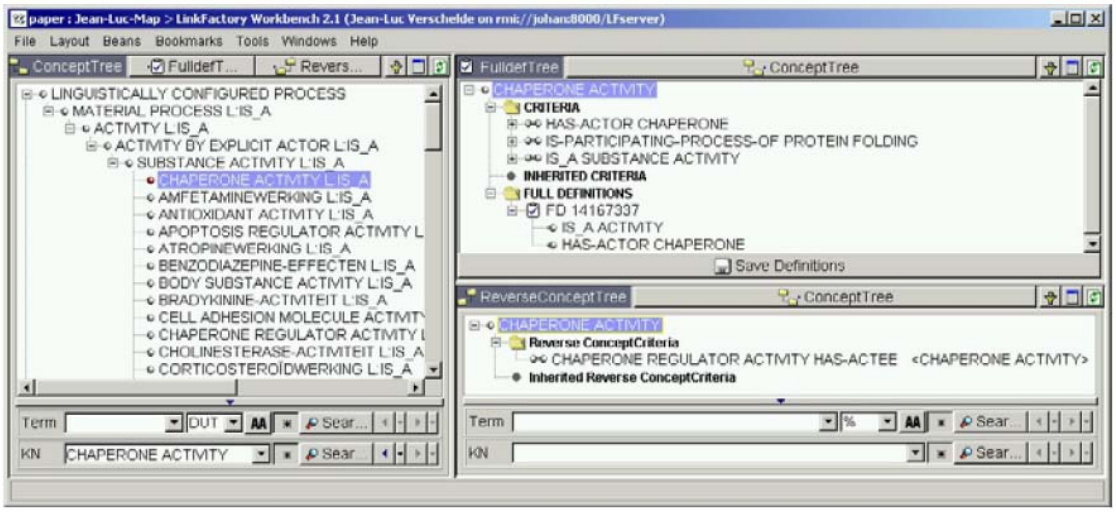


Figure 23: LinkFactory GUI

14.4.4.3 SEMEDA

SEMEDA [KOH2003] system offers semantic integration of biological databases. It is structured in three main components:

- MARGBench: offers query translation, thus enabling accessing data from distributed heterogeneous sources uniformly.
- SEMEDA: an ontology-based semantic metadatabase.
- SEMEDA-query: an ontology-based query interface.

SEMEDA allows various groups collaborating in order to construct and edit ontologies. This might be useful not only in database integration, but also in creation of general purpose ontologies, such as Gene Ontology. The collaborating groups are classified in (i) Admins, (ii) DB Provider and (iii) Everybody, having each group different permissions. SEMEDA allows them defining concepts (well defined entities with unique meaning), relations between concepts and relational algebraic properties, such as symmetry, reflexivity and transitivity. These two last properties will allow deriving the semantic entailment of concepts, and will be especially important for SEMEDA's database query interface.

SEMEDA offers a web-based interface, which was developed using JSP with oracle 8i. With it, users can query databases, guiding him in the database tables/attributes and supports the user at constructing appropriate queries. It is also possible, through the Meta DB, to browse and edit semantic database metainformation. Administrative tools are also included for performing administrative tasks, through the Admin Tools option.

14.4.4.4 Hakimpour approach

They present an approach for schema integration from different communities [HAK2001]. They propose that each group creates their own ontologies for representing concepts in the domain of their work. These ontologies will be merged based on similarity of such concepts. The final ontology, product of the fusion of all smaller ontologies, will be used to derive an integrated schema that can be used as a global schema in a federated database system.

This work pretends to solve semantic heterogeneity among different representations of data, usually due to equivalent concepts having different names. This happens very often when individual groups work on the same area of knowledge, and can be very harmful when seeking for adequate and meaningful data integration. By obtaining a global schema that integrates all local schemas, user will have a uniform and correct view of all the data. Resolving semantic heterogeneity is vital for this global schema to come out correctly, otherwise the usage of integrated data may lead to invalid results. Figure 21 shows the global schema generation approach used in this system.

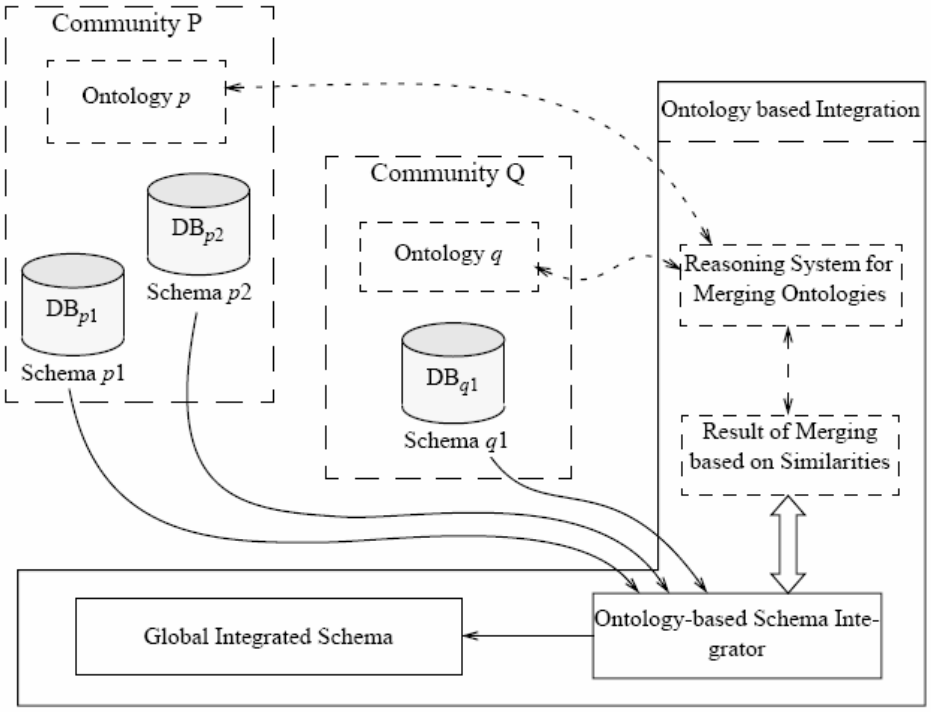


Figure 24: Global schema generation from a common ontology, result of merging local ontologies

In order to obtain a global ontology from local merging ontologies, similarities and differences among concepts must be found. Similarity relations are defined between terms found in two ontologies, based in the intensional definitions (definitions of terms by logical axioms). There are four levels of similarities between two coherent intensional definitions:

- Disjoint definitions: when the two concept or relation intentional definitions imply false. This is the level with lowest degree of similarity.
- Overlapping definitions: when the intentional definitions conjunction cannot be proven to be false.
- Specialized definitions: one of the intentional definitions is implication of the other.
- Equal definitions: both intentional definitions are equivalent.

Ontologies can be merged by means of similarities found in them. Given the level of the similarity, the merging process will be as follows:

- For equal definitions, the result is a unique intentional definition, referred to by both original terms.
- If one definition specializes an other, the similarity will be explicitly established between them.
- If one definition overlaps the other, an additional new concept or relation is declared as conjunction of both definitions.

The resulting global ontology will then be the key to build the global schema used to give users a uniform view of the data.

14.4.4.5 INFOGENMED/ONTOFUSION

INFOGENMED [PER2005a] is an information access workstation that allows biomedical professionals accessing private and public databases. Ontologies built on OWL language are used to map data sources (databases, text files, or even html pages) to virtual repositories, which are then mapped to a central virtual repository. The user is offered a graphical interface to navigate and query such repositories.

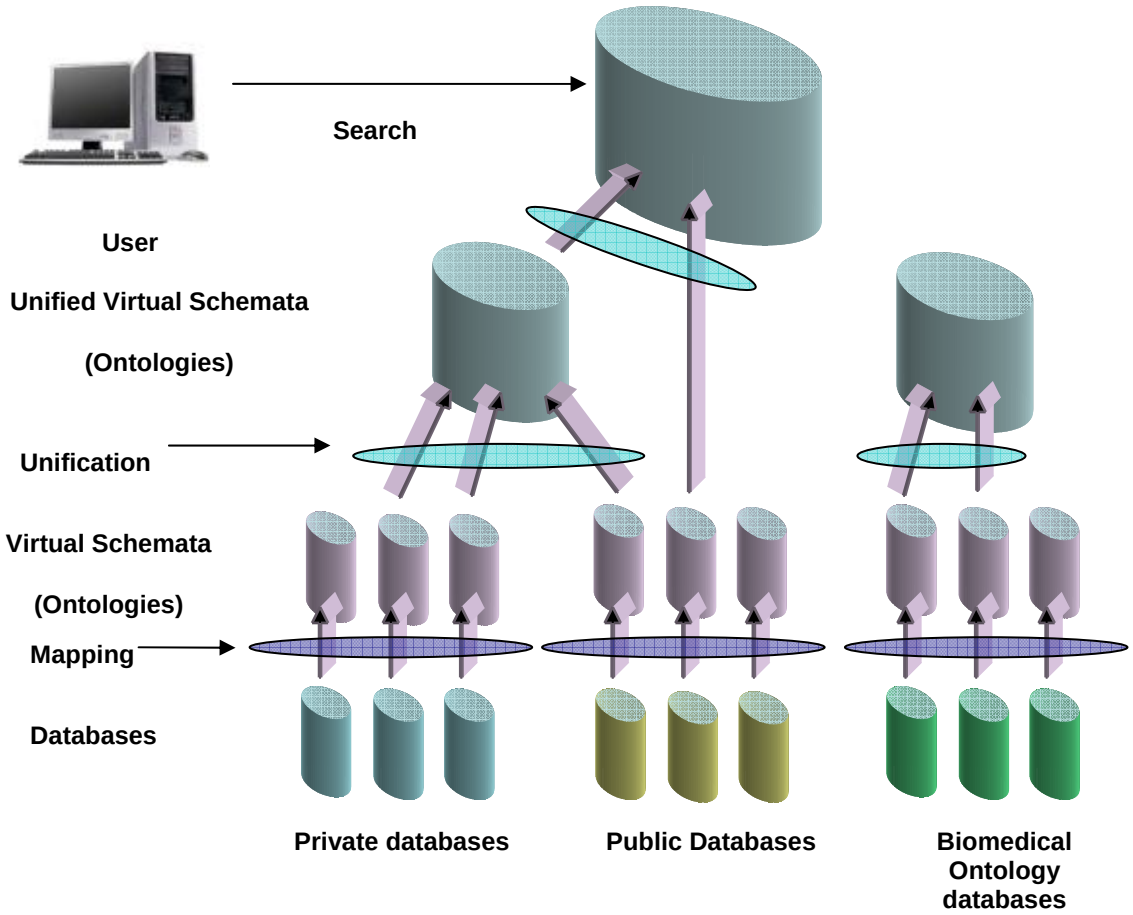


Figure 25: ONTOFUSION approach

The system was initially designed as an agent-system. The agents were in charge of solving the access problem to information sources. However, it has been recently redesigned and now it is a Web Services based system, allowing an easier access to users.

Figure 22 shows the general architecture of the INGOGENMED approach. This approach has been used in ONTOFUSION tool.

Further improvements on the system will include the integration of the OntoDataClean [PER2005b], [PER2006] tool, designed and developed by the same group responsible for ONTOFUSION. OntoDataClean tool makes use of ontologies to define the required transformations on data for cleaning and integration processes. This approach allows a more intuitive interaction with the own transformation process, allowing the specification of complex transformations on data easily. The user can query a database through this tool, specifying the cleaning ontology to be used. The system will query the database and transform the data according to the ontology prior presenting the results back to the user. The possible transformations that the ontology admits are the ones listed below:

- Missing values cleaning: this transformation allows modifying or erasing records of missing data. Missing data is defined by value ranges or specific values. The data found to be missing can be either transformed or erased.
- Format cleaning: this transformation allows modifying the data type of specific columns, which can be a requirement for subsequent integration with other data.
- Scale cleaning: this transformation allows specifying algebraic transformations on numeric data. Arithmetic operators and basic functions are allowed, as well as using previous data values as variables for calculating the resulting values.

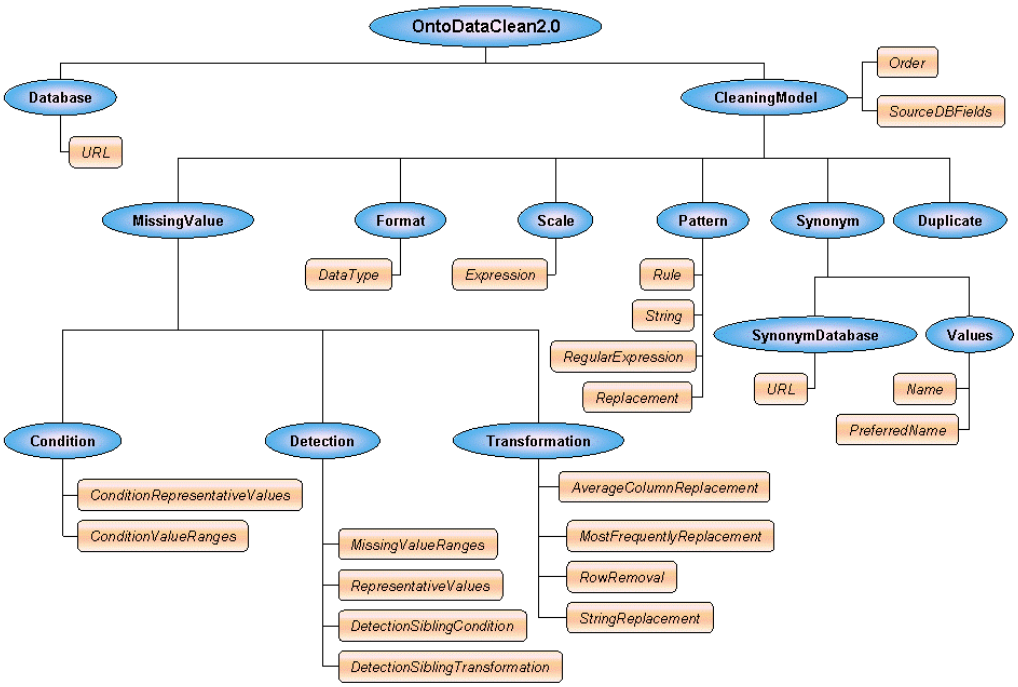


Figure 26: The cleaning ontology used by OntoDataClean

- Pattern cleaning: the pattern cleaning transformation allows modifying the pattern of string data. A powerful yet intuitive rule system is employed in order

to accomplish this task (a rule is a composition of variables and constants). The user can define rules to identify the strings to be modified, and rules to define the resulting string, allowing for example to easily specify the a transformation on a date pattern from mm-dd-yyyy to dd/mm/yy.

- Terminological inconsistencies cleaning: this transformation allows the replacement of words by preferred synonyms, given either in a specified dictionary, or explicitly.
- Duplicate cleaning: this transformation allows erasing records of data containing duplicate values in fields which are supposed to be unique (for example, employee_id).

Figure 26 shows the ontology used in OntoDataClean transformations.

OntoDataClean also includes an extension tool that analyses databases in order to find possible inconsistencies that require cleaning. This analysis is based on statistical heuristics and can provide the user with a helpful base of information before facing the specification of transformation tasks.

Further enhancements are projected for this tool, such as a wider range of transformations available, or a deeper analysis of existing inconsistencies.

14.4.5 Open Source Tools

14.4.5.1 KAON

KAON [BOZ2002] is an open source Tool suite that provides a multitude of software modules specially designed for the semantic web. It includes a persistent RDF store, an ontology store, ontology editors, etc. It has been developed as a result of a joint effort by the institute AIFB (University of Karlsruhe) and the Research Center of Information Technologies (FZI).

KAON offers an ontology management infrastructure, mainly targeted at business applications. It allows creating and managing ontologies easily and provides a framework aimed at building ontology-based applications.

KAON Reverse tool offers the possibility of mapping relational databases to ontologies, enabling both updating databases content and performing queries through the conceptualization of a database. One drawback of this tool is it incapacity of applying changes in the structure of the database to the ontology, since the whole process should be repeated and work would not be reusable.

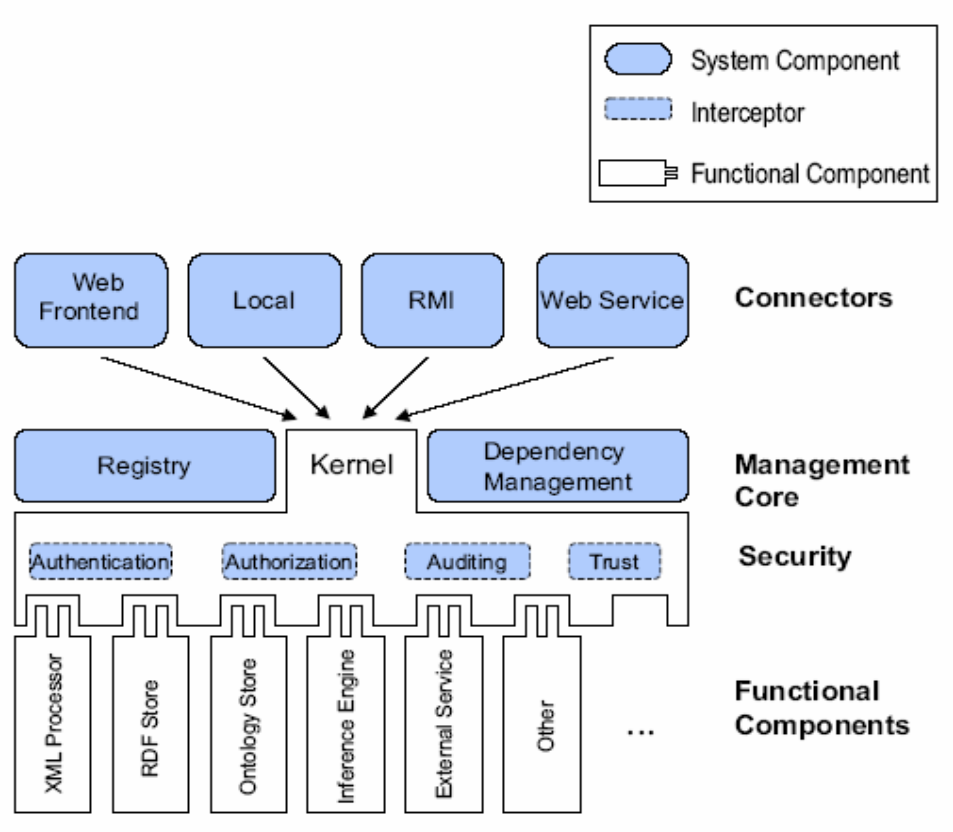


Figure 27: KAON server architecture

The kernel of this suite is the KAON SERVER, which brings all the software modules together. KAON SERVER is implemented with the Java programming language. The Java

Management Extensions (JMX) are used to manage and monitor all the resources KAON handles.

14.4.5.2 DR2 MAP

DR2 [BIZ2003] Map is a declarative and XML-based language. It allows describing mappings between relational database schemata and OWL/RDFS ontologies. With it, users can create flexible mappings of complex relational structures without having to change the existing database schema, which is achieved by applying SQL statements directly on the mapping rules.

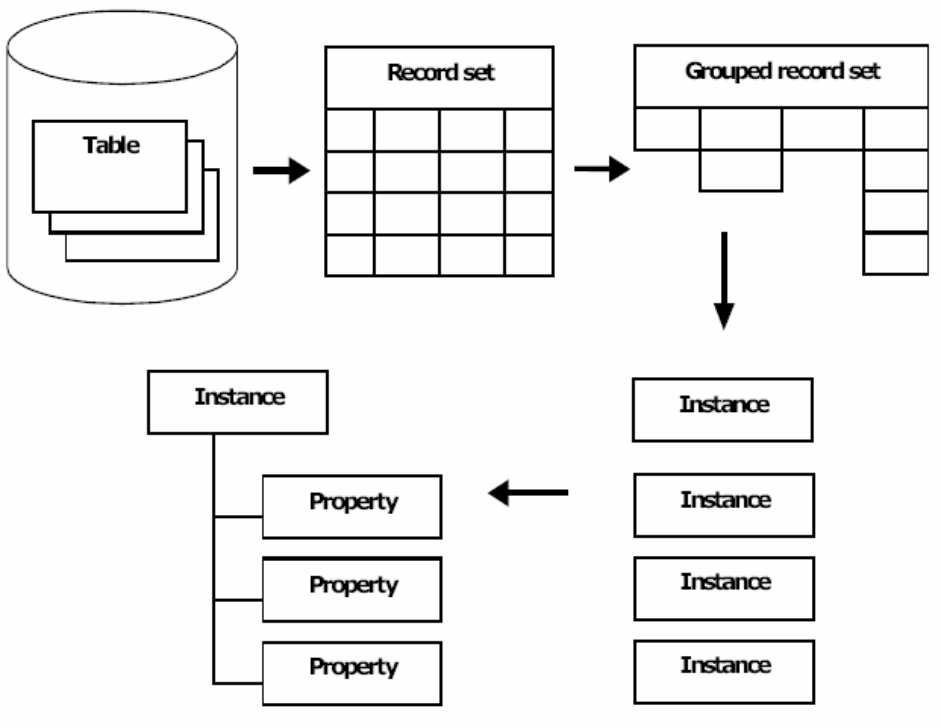


Figure 28: The DR2 mapping process

The DR2 processor is responsible for the mapping process, which is performed in four logical steps:

1. A record set is selected from the database, based on class similarity.
2. The record set is grouped according to the groupBy columns.
3. Class instances are created.
4. The record set data is mapped to instance properties.

DR2 MAP is kept as simple as possible, expressing mappings with just three elements. Figure 25 shows the mapping process used in DR2 MAP.

14.4.6 Comparison between Ontology-based Database Integration systems

In the following table it can be seen a comparison of features among several ontology-based integration systems:

	D2RMAP [13]	SEMEDA *	KAON Reverse *	INFOGENMED ONTOFUSION
Ontology Description Language	**RDF**	**RDF**	**RDFS**	**DAML+OIL**
OWL	**YES**	**No**	**No**	**YES**
Ontology Editor	**No**	**YES**	**YES**	**YES**
Ontology Graphical Browser	**No**	**No**	**YES**	**YES**
Public Database Integration	**No**	**YES**	**No**	**YES**
Physical Schema Re-design	**No**	**No**	**No**	**YES**
Virtual Schemata Unification	**No**	**No**	**No**	**YES**

14.5 References

[BIZ2003] Bizer C. “D2R MAP - A Database to RDF Mapping Language”. In Proceedings of the International World Wide Web Conference (WWW2003), Budapest, Hungary, 2003.

[BOZ2002] Bozsak E. et al. “KAON - Towards a Large Scale Semantic Web”. In K. Bauknecht, A. Min Tjoa, and G. Quirchmayr, editors, EC-Web 2002, volume 2455 of Lecture Notes in Computer Science, pages 304–313. Springer, September 2002.

[CRI1998a] Critchlow, T., Ganesh, M., and Musick, R. 1998. Meta-Data Based Mediator Generation. In *Proceedings of the 3rd IFCIS international Conference on Cooperative information Systems* (August 20 - 22, 1998). COOPIS. IEEE Computer Society, Washington, DC, 168-176. 1998

[CRI1998b] T. Critchlow, M. Ganesh, and R. Musick, "*Automatic generation of warehouse mediators using an ontology engine*," presented at the Proc. 5th KRDB Workshop, Seattle, WA, 1998.

[CRI2001] Critchlow, R. Musick, T. Slezak. Experiences applying meta-data to bioinformatics. *Inf. Sci.* 139, 1-2 (Nov. 2001), 3-17. 2001

[GRU1993] T. R. Gruber, "A Translation Approach to Portable Ontology Specifications", Knowledge Acquisition, 5(2), 199-220, 1993

[HAK2001] Hakimpour, F. and Geppert, A. 2001. Resolving semantic heterogeneity in schema integration. In *Proceedings of the international Conference on Formal ontology in information Systems - Volume 2001* (Ogunquit, Maine, USA, October 17 - 19, 2001). FOIS '01. ACM Press, New York, NY, 297-308. 2001

[KIM1996] R. Kimball. *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*, John Wiley. 1996

[KOH2003] Jacob Kohler, Stephan Philippi and Matthias Lange. SEMEDA: ontology based semantic integration of biological databases, Bioinformatics, pp 2420--2427. 2003

[PER2004] D. Perez-Rey, V. Maojo, M. Garcia-Remesal, R. Alonso-Calvo, "Biomedical Ontologies in Post-Genomic Information Systems," *bibe*, p. 207, Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'04), 2004.

[PER2005a] D. Pérez-Rey, V. Maojo, M. García-Remesal, R. Alonso-Calvo, H. Billhardt, F. Martin-Sánchez and A. Sousa, ONTOFUSION: Ontology-based integration of genomic and clinical databases, Computers in Biology and Medicine, In Press, Corrected Proof, Available online 6 September 2005

[PER2005b] D. Pérez-Rey, V. Maojo, "Nuevo modelo basado en Ontologías para el KDD en Biomedicina". TIC en Biomedicina. Colección Informática 15. ISBN 84-934497-3-3. pp. 157-176. Diciembre 2005.

[PER2006] D. Pérez-Rey, A. Anguita, V. Maojo. "OntoDataClean: Ontology-based Integration and Preprocessing of Biomedical Data". Submitted to the VII International Symposium on Biological and Medical Data Analysis (ISBMDA 06). (Submitted)

[SHE1990] A. P. Sheth e J. A. Larson. "Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases", *ACM Computing Surveys*, 22(3): pp. 183-236. 1990

[SUJ2001] W. Sujansky, Heterogeneous Database Integration in Biomedicine, Journal of Biomedical Informatics, 34, 285-298, 2001.

[VER2003] JL Verschelde, M Casella Dos Santos, T Deray, B Smith & W Ceusters Ref: Verschelde JL, Casella Dos Santos M, Deray T, Smith B and Ceusters W. Ontology-assisted database integration to support natural language processing and biomedical data-mining, Journal of Integrative Bioinformatics. 2003

[WIE1992] G. Wiederhold "Mediators in the Architecture of Future Information Systems", *IEEE Computer*, 25(3): pp. 38-49. 1992

15 Biomedical Ontologies, Terminologies and Databases

As indicated in various sections of this document, the integration strategy of ACGT is to be based on an ontology based mediation approach and a Master ACGT Ontology on Cancer.

This section presents the commonly used biomedical ontologies, terminologies and databases (OTDs) for various purposes. The OTDs have various usages within the domain of biomedicine in general and in oncology and oncology-related biology in particular. The widest service which the OTDs provide is that of a good dictionary, where different classes, terms, entities are given unique identification codes and can be used in a way that they are univocal. Arguably this is the simplest service which OTDs can provide. Ability to draw inferences, relationship among entities at various levels of granularity, existential dependence, etc. is the more advanced services which OTDs can provide. These services are used for life-science data integration, integration of Electronic Health Record data, patient status description, and drug delivery information provision in the domain of oncology. Specific features of these OTDs make them relevant for clinical practice in oncology and for oncology-related biomedical research.

15.1 Generic Medical OTDs

15.1.1 Systematized Nomenclature of Medicine – Clinical Terms (Snomed CT)

Developed by: College of American Pathologists & England and Wales National Health Service

Content: Snomed CT 9 http://www.snomed.org/snomedct/) is a generic healthcare terminology together with various relations between it's over 300,000 concepts. There are about a million descriptions of those concepts and about a million semantic links between them. The Snomed CT core content consists of:

- Concepts Table
- Descriptions Table
- Relationship Table
- History Table
- ICD Mapping

Top Classes: The main top classes consist of Clinical Finding, Procedure, Observable Entity, Body Structure, Organism, Substance, Pharmaceutical/Biologic Product, Specimen and Events.

Attributes: Snomed CT classifies attributes according to the top classes. While some attributes are used across many top classes, there are many which are characteristically used within a single top class. For example, Clinical Finding top class is associated with attributes like Severity, Onset, Course, Episodicity, Stage and so on. Similar, for Procedure,

the attributes include Procedure Site, Procedure Device, Procedure Morphology, Access and so on.

Access Rights: Snomed CT is available under license for the countries within the European Union.

Tools: Clue-5 (http://www.clininfo.co.uk/clue5/) is a CIC Lookup engine for browsing SNOMED CT and for its integration with MS Windows-based clinical applications. The Clue-5 tool provides a reference and a browser server with an API for Snomed CT integration.

Relevance to Oncology: Since Snomed CT covers the generic medical domain, there are many areas where there are overlaps with the domain of carcinomas. In particular, the classification of procedures, medications and diseases are useful. Although Snomed CT also provides an anatomical classification, the Foundational Model of Anatomy (FMA) (http://sig.biostr.washington.edu/projects/fm/AboutFM.html) seems to be more useful for carcinomas. The advantage of using Snomed, as much as possible, is that the terms are connected together and come with unique IDs. However the problems with the classifications and relationship formalisms in Snomed could lead to some limitations in inference derivation.

15.1.2 Unified Medical Language System (UMLS)

Developed by: National Library of Medicine

Content: UMLS (http://umlsinfo.nlm.nih.gov/) consists of Metathesaurus, Semantic Network, SPECIALIST Lexicon and Metamorphosys.

Metathesaurus is the vocabulary database of over a million terms dealing with the content of biomedical literature and Electronic Health Records. It consists of over 100 source vocabularies and tends to be univocal. When more than one meaning is assigned to a single vocabulary term, then both meanings of the term are represented within the Metathesaurus with the reference to specific source vocabularies. The source vocabularies integrated with the Metathesaurus includes ICD, Snomed, CPT codes, DSM, HUGO, MedDRA, NCI Thesaurus.

The Semantic Network consists of Semantic Types that provide a consistent categorization of all concepts represented in the UMLS Metathesaurus a set of Semantic Relations, which exist between Semantic Types.

The SPECIALIST Lexicon provides the lexical information needed for the SPECIALIST Natural Language Processing (NLP) System.

Access Rights: UMLS is available under license for the users within the European Union.

Tools: UMLS resources are used in applications including information retrieval, natural language processing, creation of patient and research data, and the development of enterprise-wide vocabulary services. NLM's applications include PubMed, the NLM Gateway, ClinicalTrials.gov, and the Indexing Initiative. Other examples of UMLS-enabled applications include the National Cancer Institutes Enterprise Vocabulary Services and the Agency for Healthcare Research and Quality's National Guidelines Clearinghouse and National Quality Measures Clearinghouse. UMLS knowledge sources are distributed with flexible lexical tools and the MetamorphoSys install and customization program.

Relevance to Oncology: UMLS is a conglomerate where terms from over 100 OTDs can be queried for. The Metathesaurus has been extensively used for text mining and natural

language processing in biomedical domain and thus is relevant for carcinomas. The UMLS Semantic Network and the Metathesaurus are not formalized ontologies, however, recently efforts are being made to formalize the Semantic Network in a way that inferences can be made based on it. UMLS has also been used to for mutant protein term identification from the natural text, something which helps in a semiautomatic extension of the existing mutant protein databases.

15.1.3 GALEN

Developed by: The Generalized Architecture for Languages, Encyclopedias and Nomenclatures in medicine (GALEN) project and related European Union Project Participants

Content: The GALEN project developed a Common Reference Model and a clinical terminology which can be applied to various medical domains. The GALEN project established the ontology and GRAIL formalism and demonstrated the feasibility of the concepts; GALEN-IN-USE developed the Common Reference Model (CRM) for Medical Procedures - a key element for architectures for interworking between medical records, decision support, information retrieval and natural language processing systems in healthcare. OpenGALEN was established in 1999 as a not-for-profit organisation to provide information on GALEN technologies and relevant software distributors and, in particular, to maintain and disseminate the CRM.

Access Rights: OpenGALEN (http://www.opengalen.org/) is available for free use within the European Union within the terms of its license.

Tools: Common Reference Model; GALEN Representation and Integration Language (GRAIL); Knowledge Management Environment (OpenKnoME); GALEN Case Environment

Relevance to Oncology: Similar to Snomed CT, GALEN can be embedded as a generic clinical terminology with extensions for carcinomas. GALEN is better formalized compared to the other generic medical OTDs and using GRAIL, various kinds of inferences can be derived.

15.2 Specific Medical OTDs

15.2.1 Foundational Model of Anatomy (FMA)

Developed by: Structural Informatics Group, University of Washington.

Content: FMA (http://sig.biostr.washington.edu/projects/fm/AboutFM.html) is concerned with the representation of classes and relationships necessary for the symbolic representation of the structure of the human body in a form that is understandable to humans and is also navigable by computerised systems. Specifically, the FMA is a domain ontology that represents a coherent body of explicit declarative knowledge about human anatomy. FMA has four interrelated components:

- **Anatomy taxonomy**: classifies anatomical entities according to the characteristics they share and by which they can be distinguished from one another.
- **Anatomical Structural Abstraction**: specifies the part-whole and spatial relationships that exist between the entities represented in the taxonomy

- **Anatomical Transformation Abstraction**: specifies the morphological transformation of the entities represented in the taxonomy during prenatal development and the postnatal life cycle

- **Metaknowledge**: specifies the principles, rules and definitions according to which classes and relationships in the other three components of FMA are represented.

FMA contains approximately 72,000 classes, over 115,000 terms and over 2.1 million relationship instances from 168 relationship types.

Access Rights: FMA is available for free use within the European Union within the terms of its license. A contract must be individually signed and a download access asked for.

Tools: Foundational Model Explorer is an internet based FMA browser. FMA also allows StruQL queries which provide XML as output.

Relevance to Oncology: FMA is very useful for representing anatomical entities in Relevance to Oncology. These include carcinoma staging, locations for radiotherapy and surgery, access routes for various procedures, locations for drug actions, and so on. The robust formalism allows to derivation of inferences, especially for staging of carcinomas.

15.2.2 NCI Thesaurus

Developed by: National Cancer Institute

Content: The NCI Thesaurus (http://nciterms.nci.nih.gov/NCIBrowser/Dictionary.do) is an ontology-like vocabulary that includes broad coverage of the cancer domain, including cancer related diseases, findings and abnormalities; anatomy; agents, drugs and chemicals; genes and gene products and so on. In certain areas, like cancer diseases and combination chemotherapies, it provides the most granular and consistent terminology available. It combines terminology from numerous cancer research related domains, and provides a way to integrate or link these kinds of information together through semantic relationships. The Thesaurus currently contains over 34,000 concepts, structured into 20 taxonomic trees.

Access Rights: NCI Thesaurus is available for free use within the European Union within the terms of its license.

Tools: NCI Thesaurus browser is maintained by the NCI.

Relevance to Oncology: The terminology of NCIT has been built to deal with the specific domain of carcinomas and therefore it does play an important role as a common dictionary of terms used by specialists from different domains while dealing with carcinomas. Its over-reliance on the UMLS and in particular its semantic network and some otherwise inherent problems with the classification within NCIT leads to some limitations in inference derivations; however, the NCIT does play a very useful role as a common carcinoma terminology.

15.2.3 International Classification of Diseases (ICD)

Developed by: World Health Organization

Content: ICD (http://www.who.int/classifications/icd/) is designed to promote international comparability in the collection, processing, classification, and presentation of diagnostics in health epidemiology, health management and mortality statistics. These include the analysis

of the general health situation of population groups and monitoring of the incidence and prevalence of diseases and other health problems in relation to other variables such as the characteristics and circumstances of the individuals affected. The top classes consist mainly of diseases classified according to the body system, though neoplasms, infectious diseases and injuries and poisonings have their own axes.

Access Rights: ICD is available for free use within the European Union within the terms of its license.

Tools: ICD browser is provided by the WHO. Many other browsers in different languages exist online.

Relevance to Oncology: To a large extent, ICD provides a disease classification on the basis of anatomy. Although not all the diseases within ICD are classified according to anatomy, the neoplasms are more or less classified within the anatomical partition. Thus, an ontology of carcinomas which follows the anatomical partition for classification of neoplasms and related diseases can use portions of ICD more easily than other disease classifications. However there are issues of misclassifications within ICD and also terms which do not represent a real disease. With certain modifications, integration of ICD with FMA related anatomy is possible in a way that inferences can be drawn on the basis of the anatomy ontology of FMA.

15.2.4 International Classification of Functioning, Disability and Health (ICF)

Developed by: World Health Organization

Content: ICF (http://www3.who.int/icf/icftemplate.cfm) is a classification of health and health-related domains that describe body functions and structures, activities and participation. The domains are classified from body, individual and societal perspectives. Since an individual's functioning and disability occurs in a context, ICF also includes a list of environmental factors. The top classes of ICF are: Body Functions, Body Structures, Activities and Participation and Environmental Factors. Thus ICF provides terminology not just for functions, disability and Environmental factors, but also for the body structures, although they are not formalized and detailed like other ontologies e.g. FMA.

Access Rights: ICF is available for free use within the European Union within the terms of its license.

Tools: ICD browser is provided by the WHO.

Relevance to Oncology: The classification of functioning and disability is useful to code patient status before and after therapy and also during the rehabilitation. ICF does provide a terminology which is useful for coding, however the classification is primitive and the relations between classes belonging to different axes does not exist. ICF's connection with ICD should improve the usage of both these terminologies.

15.2.5 Logical Observation Identifiers Names and Codes (LOINC)

Developed by: The Regenstrief Institute and the LOINC committee

Content: LOINC (http://www.regenstrief.org/loinc/) is a terminology primarily for laboratory results and also covers certain kinds of clinical observations. It contains over 40,000 terms out of which over 30,000 deal with the laboratory domain. The laboratory portion of the

LOINC database contains the usual categories of chemistry, haematology, serology, microbiology (including parasitology and virology), and toxicology; as well as categories for drugs, the cell counts and antibiotic susceptibility. The clinical portion of the LOINC database includes entries for vital signs, hemodynamics, intake/output, EKG, obstetric ultrasound, cardiac echo, urologic imaging, gastroendoscopic procedures, pulmonary ventilator management, selected survey instruments, and other clinical observations.

Access Rights: LOINC is available for free use within the European Union within the terms of its license.

Tools: Windows-based mapping utility called the Regenstrief LOINC Mapping Assistant (RELMA - http://www.regenstrief.org/loinc/relma/) facilitates searches through the LOINC database and assists efforts to map local codes to LOINC codes. Like the LOINC database, this program is also available for free use.

Relevance to Oncology: The LOINC database provides a terminology source which is widely used in all aspects of healthcare and thus is also widely used with the domain of carcinomas, especially in the English-speaking countries. The connection between LOINC codes and certain EHR architectures increase its usage. Most of the specific laboratory tests which are useful with respect to carcinomas are covered within LOINC. A lack of formal classification, a formal mechanism of term post-coordination and relations between various classes are known issues. However, LOINC more resembles a database than a fully fledged ontology, even though it elaborates the necessary attributes for various laboratory tests and procedures in detail.

15.2.6 Medical Subjects Headings (MeSH)

Developed by: National Library of Medicine

Content: MeSH (http://www.nlm.nih.gov/mesh/) is a controlled vocabulary thesaurus consisting of sets of terms-naming descriptors in a hierarchical structure that permits searching at various levels of specificity. The top-level classification includes: Anatomy, Organisms, Diseases, Chemicals and Drugs, Analytical, Diagnostic and Therapeutic Techniques and Equipment, Psychiatry and Psychology, Biological Sciences, and Physical Sciences. MeSH is used on MEDLINE to index bibliographic citations and author abstracts from over 4,000 journals.

Access Rights: MeSH is available for free use within the European Union within the terms of its license.

Tools: MeSH Browser provides a searchable GUI for MeSH terms. PubMed uses MeSH as its terminology to search journal articles. HONSelect is a multilingual search tool which uses MeSH to link to various healthcare-related websites.

Relevance to Oncology: MeSH is useful for the carcinoma domain due to its usage within PubMed. All major carcinoma literature is classified within PubMed and is available for retrieval using the MeSH coding. Like many other OTDs, MeSH does not claim to be a full ontology and not all its axes are as complete in terms as others. However its usage within PubMed is extensive and it has been widely embraced in many text-mining systems.

15.2.7 Medical Dictionary for Regulatory Activities (MedDRA)

Developed by: The International Conference on Harmonisation (ICH). It is owned by the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA), which

acts as trustee for the ICH steering committee. Maintained by MSSO - Maintenance and Support Services Organization.

Content: MedDRA (http://www.meddramsso.com/MSSOWeb/index.htm) is a terminology for drug and medical device side-effects and malfunctions. It emphasizes ease of use for data entry, retrieval, analysis, and display when dealing with registering, documenting, and safety monitoring of medical products. The top-level classification of MedDRA consists mainly of disorders classified according to various body systems: Respiratory disorders, Cardiac disorders, Gastrointestinal disorders, Immune system disorders, Endocrine disorders, and so on.

Access Rights: Annual subscription fee required for use within the European Union.

Tools: MedDRA browser comes with the license agreement.

Relevance to Oncology: MedDRA is used to code drug and medical device-side effects in all the medical domains and thus is also used for management of carcinomas. As far as the terminology is concerned, MedDRA tends to cover quite a generic domain similar to Snomed CT or UMLS. However the kinds of issues regarding the classification are similar to other OTDs. Not all the classes follow a classification on the basis of anatomy and this integration of MedDRA with anatomy ontologies needs reclassification and introduction of new classes. Such an effort is needed in order to improve the derivation of inferences.

15.2.8 National Drug Code Directory

Developed by: Food and Drug Administration (FDA)

Content: The Drug Listing Act of 1972 requires registered drug establishments to provide the FDA with a current list of all drugs manufactured, prepared, propagated, compounded, or processed by it for commercial distribution. Drug products are identified and reported using a unique, three-segment number, called the National Drug Code (NDC) (http://www.fda.gov/cder/ndc/) which is a universal product identifier for human drugs. FDA inputs the full NDC number and the information submitted as part of the listing process into a database known as the Drug Registration and Listing System (DRLS). Several times a year, FDA extracts some of the information from the DRLS data base for publication in the NDC Directory.

Access Rights: NDC is available for free use within the European Union within the terms of its license.

Tools: No specific publicly available tools provided.

Relevance to Oncology: The usage of NDC is mandatory for coding related to medications and this applies to all the medical domains and thus is applicable to carcinomas. Although NDC usage is mandated only within the USA, many other countries have based their requirements in lines with what is proposed by NDC. Moreover, since most of the major Hospital Information Systems and Drug Databases are NDC compliant, these codes are embedded in systems used almost everywhere in the world. NDC is not an ontology and provides a very limited set of information regarding medication and for chemotherapy agents used in carcinomas these tend to be particularly deficient. An extension of NDC is possible and is implemented within various systems.

15.2.9 Online Mendelian Inheritance in Man (OMIM)

Developed by: John Hopkins University and National Center for Biotechnology Information

Content: OMIM (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) is a catalogue of human genes and genetic disorders together with textual information and references. It illustrates the genes which have been associated with a particular disease in literature. OMIM focuses primarily on inherited or heritable, genetic diseases. It is also considered to be a phenotypic companion to the human genome project and was originally based upon the book Mendelian Inheritance in Man. Each entry is given a unique six-digit number whose first digit indicates the mode of inheritance of the gene involved

Access Rights: OMIM is available for free use within the European Union within the terms of its license.

Tools: A searchable browser with basic and advanced functions is provided. The OMIM Gene Map presents the cytogenetic locations of genes that are described in OMIM. It is a single file, presented in tabular format, listing genes from the p telomere of chromosome 1 through the q telomere of chromosome 22, followed by genes on the X and Y chromosomes. The OMIM Morbid Map is an alphabetical list of diseases described in OMIM and their corresponding cytogenetic locations.

Relevance to Oncology: The connection between gene abnormalities and diseases is useful for almost all the diseases present within OMIM. However it is especially important for hereditary diseases and carcinomas. The sheer number of genetic abnormalities associated with carcinomas is the evidence that such associations are related to the various protein and pathway abnormalities forming a part of the pathologies within carcinomas.

15.2.10 International Classification of Nursing Practice (ICNP)

Developed by: International Council of Nurses

Content: ICNP (http://www.icn.ch/icnp.htm) is a terminology which facilitates description and comparison of nursing practice. ICNP had three axes to begin with:

- Nursing phenomena (nursing diagnoses)
- Nursing actions
- Nursing outcomes

However recently the ICNP version 1 has evolved from the beta versions and consists of only one root axis – Nursing Phenomena, which in turn has seven axes – Client, Focus, Location, Judgment, Means, Time, and Action.

Access Rights: ICNP is available for free use within the European Union within the terms of its license.

Tools: Searchable browser provided by the ICN.

Relevance to Oncology: The nursing diagnoses, actions and procedures are important to every medical domain and therefore also to carcinomas. Given the large number of processes involved in management of cancer patients and the criticality of many of those procedures, a common coding system for nursing, ICNP, is highly useful. ICNP does not

claim to be an ontology and the formalism behind the classification and relationships between the classes could be further improved.

15.3 Gene Annotation OTDs

15.3.1 Gene Ontology (GO) and Gene Ontology Annotation (GOA)

Developed by: GO is developed by the Gene Ontology Consortium, of which GOA@EBI is also a part of. GOA is developed by GOA@EBI group (European Bioinformatics Institute).

Content: The GO (http://www.geneontology.org/) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. The GO project has developed three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated **biological processes**, **cellular components** and **molecular functions** in a species-independent manner. As of June 2006, GO contains 19861 terms of which 95.5% have definitions with 10690 belonging to the biological process axis, 1740 to the cellular component axis and 7431 to the molecular function. Currently, GO has only is-a and part-of relations between terms belonging to a particular axis.

GOA (http://www.ebi.ac.uk/GOA/) provides assignments of gene products to the Gene Ontology (GO) resource. UniProtKB/Swiss-Prot has joined the Gene Ontology (GO) Consortium and has adopted its standard vocabulary to characterise the activities of proteins in the UniProtKB/Swiss-Prot, UniProtKB/TrEMBL and InterPro databases. It has initiated the GOA project to provide assignments of GO terms to gene products for all organisms with completely sequenced genomes by a combination of electronic assignment and manual annotation.

Access Rights: GO and GOA are available for free use within the European Union within the terms of its license.

Tools: There are a wide plethora of tools built around GO, some by the GO consortium and many out the consortium.

- The consortium tools consist of: AmiGO, a browser allowing search for a GO term and viewing of all gene products annotated to it, or search for a gene product and viewing of all its associations and OBO-Edit, an open source, platform-independent graph-based application for viewing and editing OBO ontologies.

- Non-Consortium tools for searching and browsing GO include: CGAP GO Browser, COBrA, EP GO Browser, GeneInfoViz, GeneOntology at RZPD, GenNav, GOblet, GoFish, MGI GO Browser, QuickGO at EBI, PANDORA, TAIR Keyword Browser, Tk-GO. Tools for annotation include GeneTools, GoAnnotator, GoFigure, GoPubMed, GOtcha, HT-GO-FAT, InGOt, JAFA, Manatee and PubSearch.

- Non-consortium tools for gene expression and microarray analysis include: BiNGO, CLENCH, DAVID, EASE, eGOn v2.0, ermineJ, FatiGO, FuncAssociate, FuncExpression, GARBAN, GeneMerge, GFINDer: Genome Function, GOArray, GOdist, GOHyperGAll, GoMiner and MatchMiner, GOODIES, GOstat, GoSurfer, GO Term Finder, GOTM (Gene Ontology Tree Machine), GOToolBox, L2L, Machaon Clustering and Validation Environment, MAPPFinder, NetAffx Gene Ontology Mining Tool, Onto-Compare, Onto-Design, Onto-Express, Onto-Miner, Onto-Translate, OntoGate, Ontologizer, Ontology Traverser, Probe Explorer,

SeqExpress, SOURCE, STEM: Short Time-series Expression Miner, THEA, Avadis - gene expression analysis with GO browser and Spotfire Gene Ontology Advantage Application.

Relevance to Oncology: GO and GOA provide annotations to various gene products which are directly associated with carcinomas. The mapping of those gene products to entities within Uniprot and pathway databases and that to OMIM further close the loop by which the various functions and effects of those gene products can be queried. GO terms themselves provide a rather primitive collection of relations between the classes. However the annotations to those terms and the subsumption relationships help provide certain kinds of inferences. Despite being called an ontology, GO is far from being a formal ontology.

15.4 Protein OTDs

15.4.1 Universal Protein Resource (UniProt)

Developed by: The UniProt Consortium, which is comprised of the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR).

Content: UniProt (http://www.pir2.uniprot.org/) is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.

Access Rights: Uniprot is available for free use within the European Union within the terms of its license.

Tools: UniProt Reference Clusters (UniRef) databases combine closely related sequences into a single record to speed up searches. UniProt Archive (UniParc) is a repository with the history of all protein sequences.

Relevance to Oncology: All the sources of Uniprot provide mutant protein databases with annotation to the diseases they are associated with. The number of mutant proteins associated with carcinomas form one of the largest portion of mutant protein databases. Together with the various links to DNA and RNA databases, pathways and to biomedical literature references, Uniprot plays an important in bridging together the gap between biological and medical information related to carcinomas.

15.4.2 Structural Classification of Proteins (SCOP)

Developed by: MRC Centre for Protein Engineering, Cambridge, UK

Content: SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/) database, created by manual inspection and abetted by a battery of automated methods, provides a detailed description of the structural and evolutionary relationships between all proteins whose structure is known. It provides a broad survey of all known protein folds and detailed information about the close relatives of any particular protein. As of June 2006, SCOP has 25973 PDB Entries. The top hierarchy of SCOP includes alpha proteins, beta proteins, small proteins, multi-domain proteins, membrane and cell surface proteins, coiled coil proteins and peptides.

Availability: SCOP is available for free use within the European Union within the terms of its license.

Tools: The MRC group provides the SCOP browser, which allows browsing of the database contents including pictographic representations. There are many tools developed by other groups which use SCOP database as their major inputs. These include: Structural similarity search of SCOP using SSM, Combinatorial Extension (CE) method for structural comparison, PALI pairwise and multiple alignments of SCOP families, SUPFAM structure/sequence relationships, Structural similarity search of SCOP using 3dSearch, Structural alignment of SCOP sequences (database + server), PINTS - Patterns In Non-homologous Tertiary Structures, Sequence similarity search of SCOP using FPS, CATH structural classification, Dali structural comparison and FSSP structural classification, PDB at a Glance, and 3Dee Protein Domain Definitions.

Relevance to Oncology: SCOP provides structures of proteins on the basis of which many different protein mutant structures have been predicted. The location and depiction of various functional groups within proteins helps provide the structure-functional relationship for normal proteins which malfunction and of mutant proteins. Given the large number of mutant proteins associated with carcinomas, this information is widely applied in determination of various pathways related to carcinomas.

15.5 Pathway and Interaction OTDs

15.5.1 IntAct

Developed by: Proteomics Services Team, European Bioinformatics Institute

Content: IntAct (http://www.ebi.ac.uk/intact/index.jsp) provides a database for protein interaction data derived from literature curation or direct user submissions. IntAct also incorporates the information within interaction databases like Database of Interacting Proteins (DIP) and Biomolecular Interaction Network Database (BIND).

Access Rights: IntAct is available for free use within the European Union within the terms of its license.

Tools: The tools developed as a part of the IntAct project include:

ProViz: graph visualization system

Targets: Predicts targets for pull-down experiments

MiNe: Computes minimal connecting network for protein sets

Relevance to Oncology: The protein-protein interaction play an important role in the representation of various pathways associated with carcinomas. It not only tells about the malfunctioning related to carcinomas but also throws light on the various kinds of structural configurations of mutant proteins. Such information is also useful for drug development where agents targeting particular kinds of interactions and functional groups can be produced leading to increased efficacy and reduced adverse effects.

15.5.2 Reactome

Developed by: Cold Spring Harbor Laboratory, European Bioinformatics Institute, Gene Ontology Consortium

Content: Reactome (http://www.reactome.org/) is a curated resource of core pathways and reactions in human biology. In addition to curated human events, inferred orthologous events in 21 non-human species including mouse, rat, chicken, fugu fish, worms, fly, yeast and E.coli are also available. The main pathways represented within Reactome include:

Apoptosis, Checkpoints, Mitotic Cell Cycle, DNA Repair, DNA Replication, Electron Transport Chain, Gene Expression, Hemostasis, HIV Infection, Hs Influenza Infection, Immune System Signalling pathways, Insulin receptor mediated signalling, Integration of pathways involved in energy metabolism, Lipid metabolism, Metabolism of amino acids and related nitrogen-containing molecules, Metabolism of glucose, other sugars, and ethanol, Notch Signalling Pathway, Nucleotide metabolism, Oxidative decarboxylation of pyruvate and TCA cycle, Post-translational modification of proteins, TGF-beta signalling pathway, Transcription, Translation, mRNA Processing.

Access Rights: Reactome is available for free use within the European Union within the terms of its license.

Tools: A browsable version with explanations in detail of all the steps is provided.

Relevance to Oncology: Reactome's emphasis on pathways related to transcription and translation and to receptor communication covers a lot of turf as far as processes related to carcinomas are concerned. Pathologies behind the initiation and spread of carcinomas involve some processes which are completely absent within the normal human body. However, a majority of the processes involved in carcinomas are those which are present within the normal human body and are either abnormally regulated or over- or under-executed or take place at abnormal locations or time. The gene expression data from carcinomatous structures can be matched with respect to the expressions of various gene products.

15.5.3 Kyoto Encyclopedia of Genes and Genomes (KEGG)

Developed by: Bioinformatics Center, the Institute for Chemical Research, Kyoto University

Content: KEGG (http://www.genome.jp/kegg/) is a suite of databases and associated software, integrating the function and utility of biological systems (PATHWAY and BRITE databases), genes and proteins (GENES database), and chemical compounds and reactions (LIGAND database). The PATHWAY database covers 37,869 pathways generated from 301 reference pathways, over a million genes in their GENES database and over 14000 compounds in their LIGAND Database. The main pathways covered include:

Metabolism (Carbohydrate, Energy, Lipid, Nucleotide, Amino acid, Glycan, PK/NRP, Cofactor/vitamin, Secondary metabolite, Xenobiotics), Genetic Information Processing, Environmental Information Processing, Cellular Processes, Human Diseases and Drug Development.

Access Rights: KEGG is available for free usage within the European Union within the terms of licensing.

Tools: KEGG provides a browser which offers searching functionality together with a pictographic representation of the various pathways.

Relevance to Oncology: Like Reactome, KEGG plays an important role in oncology research. The PATHWAY database provides information relevant to the pathological processes involved in carcinoma initiation and development. Apart from the pathway-related

information, KEGG also provides information on carcinoma-relevant genes and proteins with their mutant variants.

15.6 DNA OTDs

15.6.1 Human Genome Project (HGP)

Developed by: Human Genome Project Consortium

Content: Begun formally in 1990, the U.S. Human Genome Project was a 13-year effort coordinated by the U.S. Department of Energy and the National Institutes of Health. The project originally was planned to last 15 years, but rapid technological advances accelerated the completion date to 2003. Project goals were to

- identify all the approximately 20,000-25,000 genes in human DNA,
- determine the sequences of the 3 billion chemical base pairs that make up human DNA,
- store this information in databases,
- improve tools for data analysis,
- transfer related technologies to the private sector, and
- address the ethical, legal, and social issues (ELSI) that may arise from the project.

Access Rights: HGP (http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml) sequences are available for free usage within the European Union within the terms of licensing.

Tools: Numerous projects have been spun off from the original HGP and have led to development of thousands of tools including those for analysis of gene expression, structure-function relations, transcription and translation simulators, public curation platforms like GenePoint and so on.

Relevance to Oncology: DNA structure, active zones, regulation and associated RNA transcription – all form the fundamentals of what gets coded into proteins. Given that mutant proteins and normal proteins behaving abnormally play an essential role in carcinoma initiation, development and spread, information coded within DNA molecules form the core of almost all pathologies associated with carcinomas.

15.7 RNA OTDs

15.7.1 RNA Structure Database (RNABase)

Developed by: Venkatesh Murthy, a student in the laboratory of George Rose. He can be contacted by email at venk@roselab.jhu.edu or by regular mail to Venkatesh Murthy, Johns Hopkins School of Medicine, Department of Biophysics & Biophysical Chemistry, 725 N. Wolfe St., 701 WBSB, Baltimore, MD 21205.

This information can be found at http://www.rnabase.org/about/#q10

Content: RNABase (http://www.rnabase.org/) is a database providing information regarding RNAs, especially its 3-dimensional structure in Cartesian coordinates. It uses a dedicated language to define the dihedral angles of the various RNA bonds, which provide the complete structure.

Access Rights: RNABase is available for free usage within the European Union within the terms of licensing.

Tools: A searchable browser is provided.

Relevance to Oncology: RNABase, similar to its DNA and protein counterparts, provide a possibility for structure-function comparisons. The transcription-translation process from DNA to proteins produces the catalysts and most of the participants for the biological pathways, which make them relevant for every pathological process, including those involved in the initiation and spread of carcinomas.

15.7.2 European Ribosomal RNA Database

Developed by: the University of Antwerp, Belgium. Since 2002 is maintained by the University of Ghent.

Content: European rRNA (http://www.psb.ugent.be/rRNA/) database compiles all complete or nearly complete SSU (small subunit) and LSU (large subunit) ribosomal RNA sequences. Sequences are provided in aligned format. The alignment takes into account the secondary structure information derived by comparative sequence analysis of thousands of sequences. Additional information such as literature references, taxonomy, secondary structure models and nucleotide variability maps, is also available.

Access Rights: European rRNA database is available for free usage within the European Union within the terms of licensing.

Tools: A BLASTable version of database is provided.

Relevance to Oncology: Ribosomal RNAs play an essential role in the process of translation which generates amino acid chains from the messenger RNA code. rRNAs are being actively researched for drug development in various clinical domains including oncology.

15.8 Single Nucleotide Polymorphism (SNP) OTDs

15.8.1 NIH Single Nucleotide Polymorphism Database (dbSNP)

Developed by: National Center for Biotechnology Information (NCBI)

Content: SNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) stands for "single nucleotide polymorphism". SNPs are the most common genetic variations, taking place once every 100 to 300 bases. A key aspect of research in genetics is the association of sequence variation with heritable phenotypes. It is expected that SNPs will accelerate the identification of disease genes by allowing researchers to look for associations between a disease and specific differences (SNPs) in a population. This differs from the more typical approach of

pedigree analysis which tracks transmission of a disease through a family. It is much easier to obtain DNA samples from a random set of individuals in a population than it is to obtain them from every member of a family over several generations.

Once discovered, these polymorphisms can be used by additional laboratories, using the sequence information around the polymorphism and the specific experimental conditions. The Single Nucleotide Polymorphism database (dbSNP) is a public-domain archive for a broad collection of simple genetic polymorphisms. As of June 2006, dbSNP contains over 12 million Human RefSNP clusters, over 6 million Mus musculus RefSNP clusters and over 3 million Canis familiaris and Gallus gallus clusters.

Access Rights: dbSNP is available for free usage within the European Union within the terms of licensing.

Tools: A searchable browser is provided through Entrez.

Relevance to Oncology: In the last few years, SNPs have gained a lot of importance in clinical research. The database information is compared to gene expression information of many carcinomas. Multispecies database allows comparison across different species and also make results from animal models comparable to the human case. SNPs are being widely used in chemotherapy drug development targeted against specific mutant proteins or protein complexes. Recently SNPs have also been applied for clinical research in radiotherapy.

15.8.2 Japanese Single Nucleotide Polymorphism (JSNP) Database

Developed by: JSNP is developed by Human Genome Center, Institute of Medical Science, The University of Tokyo and Japan Science and Technology Agency.

Content: JSNP (http://snp.ims.u-tokyo.ac.jp/) is the database for DNA sequence variations, polymorphic markers to investigate genes susceptible to diseases or those related to drug responsiveness. The 28th data release consists of 197,157 SNPs and 84,612 SNPs with allele frequency. SNPs will also be deposited in the public dbSNP and HGVbase (previously known as HGBASE) under the bi-directional data exchange between dbSNP and HGVbase.

Access Rights: JSNP is available for free usage within the European Union within the terms of licensing.

Tools: A searchable browser enabling gene search is provided by the developers.

Relevance to Oncology: Similar to dbSNP, JSNP database information is useful for gene expression studies and drug development.

15.9 Conclusion

From the preceeding dicusion and presentation of available Biomedical Ontologies, it becomes apparent that a significant amount of work has been devoted for the development of a large number of autonomous ontologies and/or classification systems. This forms an excellent base for ACGT's objective of drafting an integrated Master Ontology for Cancer, also including the main concepts of clinical trials, which is to act as the foundational pillar for all the "semantic" related work of the project.

16 In-Silico modelling of tumour response to therapy

16.1 Introduction

Progress in understanding cancer on the molecular level of biocomplexity has provided new weapons for fighting the disease. Nevertheless, a parallel need for satisfactorily understanding and describing cancer on the cellular and higher levels of biocomplexity, obviously depending on the molecular level, cannot be overemphasized. It is on these levels that a tumour can be definitively localised, three dimensionally imaged, geometrically and mechanically related to its neighbouring anatomical structures, spatially segmented (based on its neovasculature and subsequent metabolic activity), structurally analyzed and used as the main treatment reference by the clinician. Furthermore, an unsuccessfully treated main tumour poses a constant threat of (further) invasion and metastasis. Consequently, quantitatively understanding and virtually reproducing what is happening on *all* biocomplexity levels including the imageable tumour is a necessity.

In the last decades considerable efforts have been made in order to mathematically simulate tumour growth and tumour and normal tissue response to various therapeutic schemes. Mathematical analysis and discrete mathematics (theory of algorithms, graph theory, cellular automata, finite state machines, etc.) along with probability theory have played central roles in this process. The ultimate goal of tumour and normal tissue simulation is to contribute to the optimisation of cancer treatment by fully exploiting the individual data of the patient [STA2004], [UZU2004], and [ESC2004]. The vision is that by utilizing an “oncosimulator”, the medical doctor will be able to perform *in silico* (=on the computer) experiments corresponding to different candidate therapeutic scenarios for any cancer patient in order to facilitate and better substantiate his or her treatment decisions. Therapeutic scenarios may refer to differing radiation fractionations, differing drug administration schedules etc.

From a more theoretical point of view, computer models of tumour behaviour may also act as vehicles for the advancement of the emerging scientific and technological discipline of *In Silico* Oncology. A thorough analysis and effective simulation of the *natural phenomenon* of cancer might lead to the formulation of a number of algorithmic principles of cancer biology faintly reminiscent of Newton’s *Philosophiae Naturalis Principia Mathematica*. Obviously, a rigorous approach of this kind would have to effectively tackle both the deterministic and the stochastic character of cancer. A preliminary discussion on the subject took place in Sparta, Greece during the 1st International Advanced Research Workshop on *In Silico* Oncology, September 9-11, 2004 [STA2004], [UZU2004].

The state of the art review in this Chapter gives a short account of representative computer simulation efforts concerning tumour progression and tumour and normal tissue response to therapeutic modalities putting emphasis on molecular aspects. The focus is on the tumour growth and radiotherapy response simulation model developed by the *In Silico* Oncology Group (*IS*OG), National Technical University of Athens (www.in-silico-oncology.iccs.ntua.gr). This model may serve as a paradigm of the whole tumour simulation approach. Possible future directions are outlined at the end of the chapter.

16.2 Tumour Growth Simulation

Free tumour growth constitutes a fundamental phenomenon which takes or may take place either before tumour detection and treatment or during the intervals between subsequent treatment sessions. Two major forms of tumour progression can be distinguished: *avascular* (or prevascular) and *neovascularized* tumour growth. The first one refers to the initial development stages of a primary tumour or a micrometastasis *in vivo* or to the growth of tumour *in vitro* (e.g. a tumour spheroid in cell culture). The second one mainly refers to the progression of a clinically detectable tumour *in vivo*.

Various phenomena taking place during tumour growth have been theoretically approached by several investigators. Düchting [DUC1968] and Greenspan [GRE1976] proposed mathematical models based on control theory through which they attempted to analytically describe cancer instability. Williams & Bjerkes [WIL1972] focused on the stochasticity of abnormal clone spread. Terz et al. [TER1977] analysed the cycling and no-cycling cell populations of human solid tumours. Duechting & Vogelsaenger [DUC1981] developed a three dimensional (3D) model of spheroidal tumour growth in nutrient medium. Chen & Prewitt [CHE1982] and Balding & Mc Elwain [BAL1985] suggested mathematical representations of the neovasculaturization process. Adam & Maggelakis [ADA1990] developed a mathematical model of the diffusion regulated growth characteristics of a spherical prevascular carcinoma. Gatenby [GAT1995] investigated the competition between a tumour and the host cell population. Michelson & Leith [MIC1997] studied the possible feedback and angiogenesis mechanisms encountered in tumour growth. Retsky et al. [RET1997] developed a computer model in order to describe breast cancer metastasis. Stamatakos et al [STA1998a], [STA1998b], [STA1999] developed a Monte Carlo simulation model of avascular tumour growth and subsequently applied advanced visualisation, code parallelisation (to the limited degree allowed by tumour cells interdependence) and network techniques in order to facilitate the practical use of the model. Gödde [GOD2000] proposed a model simulating angiogenesis, vascular remodelling and haemodynamics in normal and neoplastic microcirculatory networks. Iwata [IWA2000] suggested a dynamical model for the growth and size distribution of multiple metastatic tumours. Haney et al. [HAN2001] adapted two tumour growth rate algorithms to clinical data concerning malignant gliomas. Deisboeck et al. [DEI2001], Kansal et al. [KAN2000a], [KAN2000b] and Mansury et al. [MAN2002], [MAN2003], [MAN2004a], [MAN2004b] focused on the simulation of brain tumour growth including local cell invasion progression. They introduced effective descriptive mathematical functions and special cellular automata constructs.

The *In Silico* Oncology Group (*IS*OG) avascular tumour growth simulation model [STA2001b], [STA2002], [ZAC2004a], [ZAC2004b] makes use of an appropriate cytokinetic model basically consisting of the following *states*: cell cycle phases G_1, S, G_2, Mitosis (M), G_0, Necrosis, Apoptosis and the following *state transitions*: normal cell cycling, eventual entrance to and exit from G_0, spontaneous apoptotic death, induced apoptotic death, necrotic death, cell birth and cell disappearance. A cell lying within a tumour spheroid stays in the G_0 phase for as long as its distance to glucose and oxygen supply is greater than the thickness of the outer proliferating tumour cell layer and less than the thickness of the viable tumour cell layer. A discretizing mesh of which each geometrical cell (cubic element) can be occupied either by a single tumour cell or by non tumour material (e.g. nutrient medium or normal tissue) is introduced. The tumour spheroid formation starts with the placement of either a single tumour cell at the stage of mitosis or a small tumour spheroid at the centre of the discretizing mesh. Spatial communication between cells at any angular direction is possible. The cell lysis and apoptosis products are gradually diffused towards the outer environment of the tumour. Tumour expansion is computationally achieved by shifting a cell chain from the newly occupied cubic element towards the external environment of the tumour

in a random direction. Tumour shrinkage is simulated by shifting a cell chain from the external environment of the tumour towards the cell that has to disappear in a direction defined by the cell and the centre of the mesh. Time is quantized and measured in h. The durations of the various cell states follow normal (Gaussian) distribution. The simulation can be considered a row-to-row computation of the cell algorithm for each individual cell. The outcome of a simulation run is a spatiotemporal prediction of the tumour structure and its cytokinetic activity. Further details including a successful experimental validation for the case of an EMT6 tumour spheroid are provided in the previously mentioned papers.

The *I*SOG vascularized clinical tumour growth model [STA2001a], [STA2002], [DIO2004a], [DIO2004b] although retaining certain fundamental features of the avascular (tiny) tumour growth model (e.g. cytokinetic description, Monte Carlo technique), considerably relies on the actual geometry of the imageable lesion and the spatial distribution of its metabolic activity. To this end a virtual 3D tumour reconstruction based on appropriate combinations of tomographic data collected e.g. through T1 weighted gadolinium enhanced Magnetic Resonance Imaging (MRI), Computerized Tomography (CT), Positron Emission Tomography (PET) etc. takes place before running the actual simulation. Both the spatial structure and the distribution of the metabolism/vasculature of the imageable tumour and the adjacent normal tissues are indispensable. Mechanical considerations such as the boundary conditions imposed by the skull in the case of brain tumours are made. Due to the tremendous number of tumour cells constituting a typical clinical tumour, each geometrical cell of the discretizing mesh can now be occupied by a large number of biological cells (e.g. 10^6). Biological cells contained within the same geometrical cell are clustered in equivalence classes according to the phase in which they reside at any given instant. Special consideration for the clonogenic cell density is made based on biopsy data. Parametric studies and a subsequent semi-quantitative validation have supported the applicability of the model. Interestingly, the basic philosophy of the proposed spatiotemporal gross tumour discretization strategy partly originates from the Finite Difference Time Domain (FDTD) technique which is extensively and successfully applied in a plethora of technological problems (e.g. computational electromagnetics, heat conduction etc). Once more, interdisciplinary translation of knowledge illustrates the potential of the “cross-pollination” of scientific and technological ideas.

16.3 Radiation Therapy Response Modelling and Simulation

Radiotherapy is one of the most widely applied therapeutic modalities in cancer treatment. External beam irradiation, brachytherapy, targeted radiotherapy etc. are prescribed as therapy, or for palliation or as an adjunct to surgery or chemotherapy. As the distribution of the absorbed radiation dose within the tumour and the adjacent tissues can be calculated with considerable accuracy and at the same time the mechanisms of interaction of ionizing radiation with biological tissues have been fairly elucidated, computer simulation of tumour response to radiotherapy has substantially progressed. Undoubtedly theoretical modelling of tumour response to radiotherapy lies at the heart of the treatment optimization process. To this end substantial work has been accumulated concerning mainly the response of individual tumour or normal cells to irradiation [COH1983], [FOW1997]. On the other hand models referring to the whole 3D tumour response are limited in number. The following approaches constitute representative pertinent examples.

Kocher & Treuer [KOC1995] developed a computer simulation in order to study the reoxygenation of hypoxic cells by tumour shrinkage during irradiation. Jones & Bleasdale [JON1997] modelled the influence of tumour regression and clonogen repopulation on tumour control by brachytherapy. Kocher et al. [KOC2000] simulated the cytotoxic and vascular effects of radiosurgery in solid and necrotic brain metastases. Nahum & Sanchez-

Nieto [NAH2001] developed treatment planning algorithms based on the Tumour Control Probability (TCP) that is normally used in conjunction with the notion of Normal Tissue Complication Probability (NTCP). Haney et al. [HAN2001b] mapped the therapeutic response in a patient with malignant glioma.

Stamatakos et al. [STA2001b], [STA2002] and Zacharaki et al. [ZAC2004a], [ZAC2004b] developed Monte Carlo models of the response of avascular tumours to irradiation by applying high performance computing and advanced visualisation techniques. Stamatakos et al. [STA2001a], [STA2002] and Dionysiou et al. [DIO2004a], [DIO2004b] developed simulation models of the response of large imageable dlinical tumours to radiotherapy. Clustering of tumour cells according to their proliferative status and use of the actual imaging data concerning tumour shape, metabolism and neovascularization provided a novel and promising framework for the simulation of gross tumour response to different radiotherapeutic schemes. The composite model has been semi-quantitatively validated by performing extensive parametric studies for the case of glioblastoma multiforme [STA2002], [DIO2004a], [DIO2004b], and [ANT2004a]. Large scale clinical validation and adaptation are in progress.

16.4 Chemotherapy Response Modelling and Simulation

The mechanisms of chemotherapeutic action can greatly differ among the various classes of chemotherapeutic agents. As a rule, they are more complex than those corresponding to radiotherapy response whereas at the same time the actual distribution of a drug and/or its metabolites within the tumour is difficult to predict. Nevertheless, computer simulation of chemotherapeutic schemes has also become a necessity. This is mainly due to the fact that many cancer treatment strategies rely on chemotherapy either as an exclusive modality or in combination with other techniques such as surgery and/or radiotherapy.

In the following a short account of the efforts to computer simulate tumour response to chemotherapeutic schemes is presented. Chuang [CHU1975] made specific pharmacokinetic and cell kinetic considerations for the development of mathematic models for cancer chemotherapy. Levin et al. [LEV1980] developed a heuristic model of drug delivery to malignant brain tumours. Ozawa et al. [OZA1989] performed a kinetic analysis of the cell killing effect for specific treatment cases. Jean et al. [JEA1994] introduced computer simulations to the teaching of chemotherapy. Panetta [PAN1996] developed a mathematical model of periodically pulsed chemotherapy and theoretically studied the phenomena of tumour recurrence and metastasis. Nani & Oguztereli [NAN1999] simulated the response of haematological and gynaecological cancers to chemotherapy. Iliadis & Barbolosi [ILI2000] studied the drug resistance phenomenon in cancer chemotherapy by an efficacy-toxicity mathematical model. Davis & Tannock [DAV2000] focused on the study of the repopulation of tumour cells between cycles of chemotherapy. Barbolosi & Iliadis [BAR2001] developed a pharmacokinetic-pharmacodynamic model in order to optimise drug regimens in cancer chemotherapy. Gardner [GAR2002] modelled multi-drug chemotherapy with the aim of tailoring treatment to individual patients. Ward & King [WAR2003] proposed a mathematical model of drug transport in tumour multicell spheroids and monolayer cultures.

Stamatakos et al. [STA2005] developed a spatiotemporal, patient individualised simulation model of solid tumour response to chemotherapy *in vivo* based on the actual imaging data of the patient. The *IS*OG discretizing mesh – cell clustering approach was adopted after considerable adaptations. New modules describing the pharmacokinetics and pharamacodynamics of the chemotherapeutic agent(s) were developed and appropriately integrated. The special case of glioblastoma multiforme treated by temozolomide was

considered a first application example. Good parametric behaviour of the model was demonstrated whereas clinical testing and adaptation are ongoing.

16.4.1 Simulation of Tumour Response to Other Therapeutic Modalities

Efforts to computer simulate tumour response to other treatment modalities in cellular detail have been rather scarcely recorded. As a general rule, such models tend to refer more to the physical than to the biological substrate. Two indicative examples are simulation of prostate cryoablation presented by Wojtowicz et al. [WOJ2003] and modelling of the local application of electric pulses during radiochemotherapy with tirapazamine described by Maxim et al. [MAX2004].

16.4.2 Simulation Modelling of Normal Tissue Response to Antineoplastic Interventions

Adverse effects of cancer treatment mainly refer to normal tissue response e.g. the reaction of tissues adjacent to tumour, haematopoietic system reactions etc. Toxicity (radiogenic, chemotherapeutic, etc.) plays a critical role in the therapy outcome. Therefore, it has to be carefully considered before the application of any antineoplastic scheme. Nevertheless, due to the high degree of normal tissue complexity as well as to ethical limitations, pertinent experimental knowledge on the cell level is limited. Consequently, there is scarcity of computational models (of sufficient analyticity) simulating the response of normal tissues to therapeutic interventions. Indicative examples include the NTCP model of normal tissue complications induced by radiotherapy (e.g. [NAH2001]) and the discrete state, cell cycle based radiotherapy response models described by Düchting et al. [DUC1995] and Antipas et al. [ANT2004b].

16.4.3 Integration of Molecular Networks into Tumour Behaviour Simulation Models

In order to capture the predominant mechanisms of tumour and normal tissue behaviour on *all levels* of biological complexity, tumour and affected normal tissue simulation models should inevitably incorporate information on the molecular level concerning drug-protein, radiation-gene, protein-protein, protein-DNA and other possible interactions. The spreading use of DNA and protein microarrays is providing the possibility of rationally estimating each individual cell's responsiveness to radiation therapy (e.g. through the alpha and beta parameters of the Linear Quadratic Model), to chemotherapy (e.g. through the survival fraction constant) or to other therapeutic modalities. Development of reliable molecular networks for each malignancy under consideration and for each candidate therapeutic scheme is a prerequisite for a comprehensive simulation approach [NAG2004a], [NAG2004b], [NAG2004c], Bode & Dong, [BOD2004], Pirogova et al. [PIR2002], Alcalay et al. [ALC2001]). Upon integration of critical molecular information concerning individual tumour and normal tissue cells, simulation models would logically reproduce cell-cell and cell - extracellular microenvironment interactions with considerable temporal accuracy. Subsequently, they would be capable of satisfactorily predict any candidate treatment scenario outcome. Nevertheless, inherent cancer stochasticity might still impose certain limitations in the prediction accuracy.

16.5 Future Directions

A continuous updating of oncological models based on the latest experimental and clinical data is essential to both cancer understanding and individualized treatment optimization. This implies that any practical simulation model should always be amenable to considerable extensions and/or modifications in order to incorporate new knowledge as well as bright ideas emerging at an astonishingly fast rate (e.g. Simpson et al. [SIM2004]). Parametrical, experimental and clinical validation as well as adaptation of the models should necessarily follow each eventual modification process [STA2006]. Compatibility with current imaging and molecular data formats and at the same time exploitation of the constantly increasing potential of computer technology in terms of both processing rate and memory should be technical considerations of high priority.

Concerning the range of applicability of tumour simulation models, the following areas might also be targeted in the near future:

- Identification of potential tumour vulnerabilities suggesting new therapeutic strategies in the generic (research) context;
- education of medical doctors, life scientists, researchers and interested patients by virtual reality demonstrations of the likely response of an arbitrary tumour to different candidate treatment schemes.

16.6 References

[ADA1990] Adam, J. A. & Maggelakis, S. A. 1990, “Diffusion regulated growth characteristics of a spherical prevascular carcinoma,” *Bull. Math. Biol.*, vol. 52, pp. 549–582.

[ALC2001] Alcalay, M., Orleth, A., Sebastiani, C., Meani, N., Chiaradonna, F., Casciari, C., Sciurpi, M.T., Gelmetti, V., Riganelli, D., Minucci, S., Fagioli, M., and Pelicci, P.G. 2001, “Common themes in the pathogenesis of acute myeloid leukaemia”, *Oncogene*, vol. 20, pp. 5680-5694.

[ANT2004a] Antipas, V.P., Stamatakos, G. S., Uzunoglu, N.K., Dionysiou, D. D. & Dale, R. G. 2004 a, “A spatio-temporal simulation model of the response of solid tumours to radiotherapy *in vivo*: parametric validation concerning oxygen enhancement ratio and cell cycle duration”, *Phys. Med. Biol.*, vol. 49, pp. 1485–1504.

[ANT2004b] Antipas, V. P., Stamatakos, G.S., Uzunoglu, N.K., Kouloulias, V.E. 2004 b, “Towards a spatiotemporal simulation of the in vivo response of normal tissues to radiotherapy”, in *Book of Short Communications*, *First International Advanced Research Workshop on In Silico Oncology: Advances and Challenges, Sparta, Greece, September 9-10*, *2004*, ed. N. Uzunoglu, G.Stamatakos, D. Givol, Institute of Communications and Computer Systems, National Technical University of Athens, Athens, Greece, 2004, pp.65-67.

[BAL1985] Balding, D. & Mc Elwain, D.L.S. 1985, “A mathematical model of tumor-induced capillary growth,” *J.theor. Biol.*, vol. 114, pp.53-73.

[BAR2001] Barbolosi, D. & Iliadis A. 2001, “Optimizing drug regimens in cancer chemotherapy: a simulation study using a PK-PD model,” *Comput. Biol. Med.*, vol. 31, pp. 157-172.

[BOD2004] Bode, A. & Dong, Z. 2004, “Post-translational modification of p53 in tumorigenesis”, *Nature Rev. Cancer*, vol. 4, pp. 793-805.

[CHE1982] Chen I.I.H. & Prewitt, R.L. 1982, “A mathematical representation for vessel network,”

J. theor. Biol., vol. 97, pp.211-219.

[CHU1975] Chuang, S. 1975, "Mathematic models for cancer chemotherapy: pharmacokinetic and cell kinetic considerations," *Cancer Chemother. Rep.* vol. 59, no. 4, pp. 827-42.

[COH1983] Cohen, L. 1983, *Biophysical Models in Radiation Oncology*, CRC Press, Boca Raton.

[DAV2000] Davis, J. & Tannock, I. F. 2000, "Repopulation of tumour cells between cycles of chemotherapy: a neglected factor," *The Lancet Oncology*, vol. 1, pp.86-93.

[DEI2001] Deisboeck, T.S., Berens, M.E., Kansal, A.R., Torquato, S., Stemmer-Rachamimov, A.O., & Chiocca, E.A. 2001 "Pattern of self-organization in tumor systems: complex growth dynamics in a novel brain tumor spheroid model", *Cell Prolif.* vol. 34, pp. 115-134.

[DIO2004a] Dionysiou, D.D. Stamatakos, G.S., Uzunoglu, N.K., Nikita, K.S., Marioli, A. A 2004 a, "Four dimensional in vivo model of tumour response to radiotherapy: parametric validation considering radiosensitivity, genetic profile and fractionation", *J. theor. Biol.*, vol. 230, pp.1-20.

[DIO2004b] Dionysiou, D. 2004 b, "Computer simulation of *in vivo* tumour growth and response to radiotherapeutic schemes. Biological optimization of radiation therapy by *"in silico"* experimentation.", PhD thesis (*in Greek*), Dept. of Electrical and Computer Engineering, National Technical University of Athens, 2004.

[DUC1968] Düchting, W. 1968, "Krebs, ein instabiler Regelkreis, Versuch einer Systemanalyse", *Kybernetik*, Band 5, Heft 2, pp. 70 – 77.

[DUC1981] Düchting, W. & Vogelsaenger, T. 1981, "Three-dimensional pattern generation applied to spheroidal tumor growth in a nutrient medium," *Int. J. Biomed. Comput.*, vol. 12, no. 5, pp. 377-392.

[DUC1995] Düchting, W., Ulmer, W., Ginsberg, T., Kikhounga-N Got, O. & Saile, C. 1995, "Radiogenic Responses of Normal Cells Induced by Fractionated Irradiation – a Simulation Study", *Strahlenther. Onkol.* vol. 171, pp.460-467.

[ESC2004] von Eschenbach, A.C. 2004, "A vision for the National Cancer Program in the United States", *Nature Rev. Cancer*, vol.4, pp.820-828.

[FOW1997] Fowler, J.F. 1997, "Review of radiobiological models for improving cancer treatment," in *Modelling in Clinical Radiobiology*, ed. K.Baier & D. Baltas, Freiburg, Germany, Albert-Ludwigs-University, Freiburg Oncology Series, Monograph No.2, ch.1, pp.1-14.

[GAR2002] Gardner, S.N. 2002, "Modeling multi-drug chemotherapy: tailoring treatment to individuals," *J. theor. Biol.*, vol. 214, pp.181-207.

[GAT1995] Gatenby, R.A. 1995, "Models of tumor-host interaction as competing populations: implications for tumor biology and treatment," *J.theor. Biol.*, vol. 176, pp.447-455.

[GOD2000] Gödde, R., Düchting, W. & Kurz, H. 2000, "Simulation of Angiogenesis, Vascular Remodelling and Haemodynamics in Normal and Neoplastic Microcirculatory Networks", *Annals of Anatomy*, vol. 182 Supplement, pp. 9-10.

[GRE1976] Greenspan, H.P. 1976, "On the growth and stability of cell cultures and solid tumors," *J.thor. Biol.*, vol. 56, pp.229-242.

[HAN2001] Haney, S., Thompson, P. M., Cloughesy, T. F., Alger, J. R., Toga, 2001 a, "A. W.

Tracking tumor growth rates in patients with malignant gliomas. A test of two algorithms." *American Journal of Neuroradiology* vol. 22, pp. 73-82.

[HAN2001b] Haney, S., Thompson, P. M., Cloughesy, T. F., Alger, J.R., Frew, A., Torres-Trejo, A., Mazziotta, J.C, Toga, A. W. 2001 b, "Mapping therapeutic response in a patient with malignant glioma", *Journal of Computer Assisted Tomography*, vol. 25, pp.529-536.

[ILI2000] Iliadis, A. & Barbolosi, D. 2000, "Optimizing drug resistance in cancer chemotherapy by an efficacy-toxicity mathematical model," *Comput. Biomed. Res.*, vol. 33, pp.211-226.

[IWA2000] Iwata, K., Kawasaki, K. & Shigesada, N. 2000 b, "A dynamical model for the growth and size distribution of multiple metastatic tumors", *J. theor. Biol.*, vol. 201, pp.177-186.

[JEA1994] Jean, Y., De Traversay, J., Lemieux, 1994, P. "Teaching cancer chemotherapy by means of a computer simulation," *Int. J. Biomed. Comput.*, vol. 36, pp.273-280.

[JON1997] Jones, B. and Bleasdale, C. 1997, "Influence of Tumour Regression and Clonogen Repopulation on Tumour Control by Brachytherapy," in *Modelling in Clinical Radiobiology*, eds. K.Baier and D. Baltas, Freiburg, Germany, Albert-Ludwigs-University, Freiburg Oncology Series, Monograph No.2, ch.14, pp.116-126.

[KAN2000a] Kansal, A. R., Torquato, S., Harsh, G. R., Chiocca, E. A. & Deisboeck, T. S. 2000 a, "Cellular Automaton of Idealized Brain Tumor Growth Dynamics", *BioSystems*, vol. 55, pp. 119-127.

[KAN2000b] Kansal, A.R. Torquato, S. Harsh, G.R., Chiocca, E.A. & Deisboeck, T.S. 2000 b, "Simulated brain tumor growth dynamics using a three-dimensional cellular automaton," *J.theor. Biol.,* vol. 203, pp.367-382.

[KOC1995] Kocher, M. & Treuer, H. 1995, "Reoxygenation of hypoxic cells by tumor shrinkage during irradiation. A computer simulation.", *Strahlenther. Onkol.,* vol. 171, pp.219-230.

[KOC2000] Kocher, M. Treuer, H., Voges, J., Hoevels, M., Sturm, V.,.Mueller, R.P. 2000, "Computer simulation of cytotoxic and vascular effects of radiosurgery in solid and necrotic brain metastases," *Radiother. Oncol.,* vol. 54, pp.149-156.

[LEV1980] Levin, V.A. Patlak, C.S. Landahl, H.D. 1980, "Heuristic modeling of drug delivery to malignant brain tumors," *J. Pharmacokinet. Biopharm.*, vol. 8, pp.257-296.

[MAN2002] Mansury, Y., Kimura, M., Lobo, J., & Deisboeck, T.S. 2002, "Emerging patterns in tumor systems: simulating the dynamics of multicellular clusters with an agent-based spatial agglomeration model", *J. Theor. Biol.*, vol. 219, pp. 343-370.

[MAN2003] Mansury, Y. & Deisboeck, T.S. 2003, "The impact of 'search precision' in an agent-based tumor model", *J. Theor. Biol.*, vol. 224, pp. 325-337.

[MAN2004a] Mansury, Y. & Deisboeck, T.S. 2004 a, "Simulating 'structure-function' patterns of malignant brain tumors", *Physica A*, vol. 331, pp. 219-232.

[MAN2004b] Mansury, Y., Athale, C., Gregor, B. F. & Deisboeck, T. S. 2004 b, "Modeling malignant brain tumors with a novel spatio-temporal agent-based simulation framework", in *Book of Short Communications*, *First International Advanced Research Workshop on In Silico Oncology: Advances and Challenges, Sparta, Greece, September 9-10*, *2004*, ed. N. Uzunoglu, G.Stamatakos, D. Givol, Institute of Communications and Computer Systems, National Technical University of Athens,

Athens, Greece, 2004, pp.45-47.

[MAX2004] Maxim,P.G., Carson,J.J.L., Ning,S., Knox,S.J., Boyer, A.L., P. Hsu,C.P., Benaron,D.A. & Walleczek, J. 2004, "Enhanced effectiveness of radiochemotherapy with tirapazamine by local application of electric pulses to tumors", *Radiation Research*, vol. 162, pp.185-193.

[MIC1997] Michelson, S. & Leith, J.T. 1997, "Possible feedback and angiogenesis in tumor growth control," *Bull. Math. Biol.*, vol. 59, pp.233-254.

[NAG2004a] Nagl, S. B. 2004 a, "Modelling complex cellular systems for post-genomic biomedicine" in *Computation in Cells and Tissues: Perspectives and Tools of Thought*, eds Paton R. C. et al., Springer Verlag.

[NAG2004b] Nagl, S. B. & Patel, M. 2004 b, "MicroCore: Mapping genome expression to cell pathways and networks", *Comparative and Functional Genomics*, vol. 5, pp. 75-78.

[NAG2004c] Nagl,S. 2004 c, "Modelling of cancer gene-signal systems" *Book of Short Communications*, *First International Advanced Research Workshop on In Silico Oncology: Advances and Challenges, Sparta, Greece, September 9-10*, *2004*, ed. N. Uzunoglu, G.Stamatakos, D. Givol, Institute of Communications and Computer Systems, National Technical University of Athens, Athens, Greece, 2004, pp.23-24.

[NAH2001] Nahum, A. & Sanchez-Nieto, B. 2001, "Tumour control probability modelling: basic principles and applications in treatment planning", *Physica Medica,* vol. 17 (xvii), supl. 2, pp.13-23.

[NAN1999] Nani, F.K. & Oguztereli, M.N. 1999 "Modelling and simulation of chemotherapy of haematological and gynaecological cancers," *IMA J. Math. Appl. Med. Biol.*, vol. 16, pp.39-91.

[OZA1989] Ozawa, S. Sugiyama, Y. Mitsuhashi, J. & Inaba, M. 1989 "Kinetic analysis of cell killing effect induced by cytosine arabinoside and cisplatin in relation to cell cycle phase specificity in human colon cancer and chinese hamster cells" *Cancer Res.*, vol. 49, pp. 3823-3828.

[PAN1996] Panetta, J.C. 1996, "A mathematical model of periodically pulsed chemotherapy: tumor recurrence and metastasis in a competitive environment" , *Bull.Mat. Biol.*, vol. 58, pp.425-447.

[PIR2002] Pirogova, E.,Fang, Q. Akay, M., Cosic, I. 2002, "Investigation of the Structural and Functional Relationships of Oncogene Proteins", *Proc of IEEE*, vol. 90, pp. 1859-1867.

[RET1997] Retsky, M. W., Demicheli, R., Swartzendruber, D., E., Bame, P. D., Wardwell, R.H., Bonadonna, G., Speer, J. F., Valagussa, P. 1997, "Computer simulation of a breast cancer metastasis model", *Breast Cancer Res. Treat.*, vol. 45, no. 2, pp. 193-202.

[SIM2004] Simpson, M., Cox, C., Peterson, G., Slayer, G. 2004, "Engineering in the biological substrate: information processing in genetic circuits", *Proc. IEEE*, vol. 92, no.5, pp.848-863.

[STA1998a] Stamatakos, G., Uzunoglu, N., Delibasis, K., Makropoulou M., Mouravliansky, N. & Marsh, A. 1998 a, "A simplified simulation model and virtual reality visualization of tumor growth *in vitro"*, *Future Generation Comput. Syst.*, vol. 14, pp. 79–89

[STA1998b] Stamatakos, G.S., Uzunoglu, N.K., Delibasis, K., Makropoulou, M., Mouravliansky, N., Marsh, A. 1998 b, "Coupling parallel computing and the WWW to visualize a simplified simulation of tumor growth in vitro," *Proc. International Conference on*

Parallel and distributed Processing Techniques and Applications, PDTA'98, Las Vegas, USA, ed.H.R.Arabnia, CSREA Press, pp.526-533.

[STA1999] Stamatakos, G., Zacharaki, E., Mouravliansky, N., Delibasis, K., Nikita, K., Uzunoglu, N., A.Marsh, 1999, "Using Web technologies and meta-computing to visualize a simplified simulation model of tumor growth in vitro" in *High-Performance Computing and Networking*, ed. P.Sloot, M.Bubak, A.Hoekstra, B. Hertzberger Lecture Notes in Computer Science vol. 1593, Springer, Berlin, pp.973-982.

[STA2001a] Stamatakos, G., Dionysiou, D., Nikita, K., Zamboglou, N., Baltas, D., Pissakas, G. & Uzunoglu, N. 2001 a, '*In vivo* tumor growth and response to radiation therapy: A novel algorithmic description', *Int. J.Radiat. Oncol. Biol. Phys*., vol. 51, no. 3, Sup. 1, p. 240.

[STA2001b] Stamatakos, G., Zacharaki, E., Makropoulou, M., Mouravliansky, N. Marsh, A., Nikita, K. & Uzunoglu, N. 2001 b. "Modeling tumor growth and irradiation response in vitro - a combination of high-performance computing and web based technologies including VRML visualization" *IEEE Trans. Inform. Technology Biomedicine*, vol. 5, no 4, pp.279-289.

[STA2002] Stamatakos, G., Dionysiou, D., Zacharaki, E., Mouravliansky, N., Nikita, K. & Uzunoglu, N. 2002, "*In Silico* Radiation Oncology: Combining Novel Simulation Algorithms with Current Visualization Techniques", *Proc. IEEE, Special Issue on "Bioinformatics: Advances and Challenges*, Vol.90, No.11, pp.1764-1777 *(invited paper)*

[STA2004] Stamatakos, G. S. 2004 a, "The ISOG/ICCS/NTUA in silico model of in vivo tumour and normal tissue response to radiation therapy and chemotherapy: The modules of the model", in Book of Short Communications, First International Advanced Research Workshop on In Silico Oncology: Advances and Challenges, Sparta, Greece, September 9-10, 2004, ed. N. Uzunoglu, G.Stamatakos, D. Givol, Institute of Communications and Computer Systems, National Technical University of Athens, Athens, Greece, 2004,pp.51-53

[STA2005] G.S.Stamatakos GS, V.P. Antipas VP, N.K. Uzunoglu, "Simulating chemotherapeutic schemes in the individualized treatment context: The paradigm of glioblastoma multiforme treated by temozolomide in vivo." *Comput Biol Med.* 2005 Oct 2;

[STA2006] G. S. Stamatakos, V.P. Antipas, N. K. Uzunoglu, R. G. Dale, "A four dimensional computer simulation model of the *in vivo* response to radiotherapy of glioblastoma multiforme: studies on the effect of clonogenic cell density." *British Journal of Radiology,* 2006, vol. 79, 389-400

[TER1977] Terz, J.J., Lawrence W. Jr. & Cox, B. 1977, "Analysis of the cycling and noncycling cell population of human solid tumors", *Cancer*, vol.40, pp.1462-1470.

[UZU2004] Uzunoglu, N. K. 2004, "The *IS*OG/ICCS/NTUA *in silico* model of *in vivo* tumour and normal tissue response to radiation therapy and chemotherapy: the principles", in *Book of Short Communications*, *First International Advanced Research Workshop on In Silico Oncology: Advances and Challenges, Sparta, Greece, September 9-10*, *2004*, ed. N. Uzunoglu, G.Stamatakos, D. Givol, Institute of Communications and Computer Systems, National Technical University of Athens, Athens, Greece, 2004,pp.48-50.

[WAR2003] Ward, J.P. & King, J.R. 2003, "Mathematical modeling of drug transport in tumour multicell spheroids and monolayer cultures," *Mat. Biosci.,* vol. 181, pp.177-207.

[WIL1972] Williams, T.& Bjerknes, R. 1972, "Stochastic model for abnormal clone spread

through epithelial basal layer", *Nature*, vol. 236, pp. 19-21.

[WOJ2003] Wojtowicz, A., Selman, S., Jankun, J. 2003, " Computer simulation of prostate cryoablation--fast and accurate approximation of the exact solution", *Comput Aided Surg.*, vol. 8, np. 2, pp. 91-97.

[ZAC2004a] Zacharaki, E. I., Stamatakos, G. S., Nikita, K. S. & Uzunoglu, N. K. 2004 a, "Simulating growth dynamics and radiation response of avascular tumour spheroids – Model validation in the case of an EMT6/Ro multicellular spheroid", *Computer Methods and Programs in Biomedicine*, vol. 76, pp.193-206.

[ZAC2004b] Zacharaki, E. 2004 b, "Development of imaging data registration algorithms and computer simulation models of malignant tumour behaviour aiming at supporting clinical decisions in radio-oncology", PhD thesis (*in Greek*), Dept. of Electrical and Computer Engineering, National Technical University of Athens, 2004.

17 Data Mining and Knowledge Discovery

A key-technology for addressing the some of the challenges in BioMedical Informatics in general and the ACGT workplan in specific relates to *data mining* systems, methods and tools. Data mining is a step in the process of generating knowledge in databases. It includes techniques for query databases, on-line analytical processing, and machine-learning algorithms. In the medical area, many applications have been created for decision support to address issues such as image and signal analysis and outlining clinical prognoses for patient conditions. In biology, efforts have been centered on research issues such as the prediction of protein structures and drug studies.

Both types of predictive exercises present considerable challenges for future research. Text mining is a discipline that aims to extract data, information or knowledge from texts. Finding information in biomedical databases using text mining and information-retrieval techniques is expected to leverage a substantial amount of biomedical information that has escaped analysis until now.

Some of the key biomedical tasks to be tackled with data-mining follow.

- *Genome Database Mining.* Genome database mining is an emerging technology that is based on extracting useful information from genome databases. One of the main tasks is the computational annotation of genomes, that consists of two sequential processes [BOR1998; ROU1999]:

    - *Structural annotation*: refers to the identification of hypothetical genes termed open reading frames (ORFs) in a DNA sequence using computational gene discovery algorithms. *Functional annotation:* refers to the assignment of functions to the predicted genes using sequence similarity searches against other genes of known function.

- *Computational/Mining for Gene Discovery*. The finding of genes on a genome is a complex task. The regions that code for proteins (exons) are only a tiny fraction of the genome. These regions can be predicted making use of the biological properties and the particular statistical composition that characterize these regions. Computational gene discovery techniques are able to find these dispersed coding exons in a sequence and to provide the best tentative gene models. Exon recognition algorithms exhibit performance tradeoffs between increasing sensitivity - ability to detect true positives, and decreasing specificity - ability to exclude false positives (http://www.nslij-genetics.org/gene/).

- *Sequence Similarity Searching*. Sequence similarity searching is an important methodology in computational molecular biology. Initial clues to understanding the structure or function of a molecular sequence arise from similarity to other molecules that have been previously studied. Sequence database searches reveal biologically significant sequence relationships and suggest future investigation strategies (http://www.ebi.ac.uk/Tools/similarity.html). Sequence similarity searches is mainly exploited for:

- *Comparative Genomics* (http://www.ornl.gov/sci/techresources/Human_Genome/faq/compgen.shtml) - the analysis and comparison of genomes from different species. The purpose is to gain a better understanding of how species have evolved and to determine the function of genes and noncoding regions of the genome. Comparative genomics involves the use of computer programs that can line up multiple genomes and look for regions of similarity among them. Some of these sequence-similarity tools are accessible to the public over the Internet. One of the most widely used is BLAST (http://www.ncbi.nlm.nih.gov/BLAST/), which is available from the National Center for Biotechnology Information. BLAST is a set of programs designed to perform similarity searches on all available sequence data.

- *Gene Expression Mining*. Is defined as the use of quantitative messenger RNA (mRNA)-level measurements of gene expression in order to characterize biological processes and elucidate the mechanisms of gene transcription. Changes in gene expression under the influence of drug or disease perturbations can be studied. The identification of differential gene expression associated with biological processes is a central research problem in molecular genetics. High throughput gene expression assays enable the simultaneous monitoring of thousands of genes in parallel and generate vast amounts of gene expression data. The large-scale investigation of gene expression attaches functional activity to structural genetic maps and therefore is an essential milestone in the paradigm shift from static structural genomics to dynamic functional genomics. Gene expression database mining is used to identify intrinsic patterns and relationships in gene expression data. The identification of patterns in complex gene expression datasets provides two benefits: (a) Generation of insight into gene transcription conditions; (b) Characterization of multiple gene expression profiles in complex biological processes, e.g. pathological states. Data visualization is used to display snapshots of cluster analysis results generated from large gene expression datasets (http://www.computational-genomics.net/genomics_9.html).

- *Proteomics and Data Mining*. The study of the proteome is important because proteins represent the actual functional molecules in the cell. Proteomics covers a number of different aspects of protein function, including the following: (a) Structural proteomics - the large-scale analysis of protein structures; (b) Expression proteomics - the large-scale analysis of protein expression, this can help to identify the main proteins found in a particular sample and proteins differentially expressed in related samples, such as diseased vs healthy tissue; (c) Interaction proteomics - the large-scale analysis of protein interactions; the characterization of protein-protein interactions helps to determine protein functions and can also show how proteins assemble in larger complexes (http://www.wellcome.ac.uk/en/genome/thegenome/ hg03b002.html).

- *Metabolomics and Data Mining*. Proteomic analysis methods such as mass spectrometry allow the abundance and distribution of many proteins to be determined simultaneously. Mass-spectral raw data are pre-processed in order to convert them to a form usable by data mining algorithms. Here the mining task is that of feature extraction and classification, i.e., peak detection and peak calibration, and then clustering - clusters of (detected) mass-spectra peaks become the extracted features. Once the mass spectral data have been pre-processed and their features extracted one proceed to biomarker discovery including support vector machines, neural networks, decision trees and more [HIL2004].

In order to advance data mining within the BioMedical Informatics (BMI) context, objectives and goals, special R&D efforts should be forwarded in the utilization of main data mining standards and libraries:

- ***PMML*** (Predictive Model Markup Language, from the Data Mining Group – DMG; http://www.dmg.org/) – a standard XML mark up language to describe statistical and data mining tools, and

- ***Weka*** – an interoperable suite of data mining components and tools (http://www.cs.waikato.ac.nz/~ml/weka/).

- ***R-package****: An integrated environment for data-mining.* Over the last decade, the ***R*** *open source statistical package* (www.r-project.org) has quickly become the platform of choice for statistical research as well as many applied statistics projects. R is a widely used open source language and an environment for statistical computing and graphics. It has mechanisms to interact directly with software that has been written in many different languages. These tools allow users to incorporate modules based on other work. So adopting R does not exclude other development environments. R can, in those cases, provide a *glue* or connectivity linking. This solution has significant advantages:

    - R is quickly becoming a de facto standard in statistical computing and is already widely used in biostatistics and health care.

    - R is mature, state of the art, and extremely comprehensive.

    - R provides: Advanced data mining methods; Comprehensive packages for survival analysis (Kaplan-Meier, Cox proportional hazards, Bayesian survival, etc) essential for cancer research; Standard hypothesis testing; Tools for linear and non-linear regression, discriminant analysis, SVM, clustering (k-means, SOMs etc); Extensive visualization capabilities; Special packages (Bioconductor, www.bioconductor.org) for the analysis of genomic data.

    - R is quite well-connected and reads text-files, databases (ODBC, Excel, and Access that are heavily used in health care, and mysql, postgres, Oracle, and others), and has Java connectivity.

    - It is well documented, extensible, and offers a programming language.

Since R is open source, and available free of charge, the consortium does not have to rebuild already available components (hypothesis testing, basic survival analysis, basic statistical visualization), but can build on proven technology, and focus on advancing the state of the art.

17.1 Mining the Biomedical Literature

Biomedical literature data mining is the process of identifying and extracting valid, novel and useful nuggets of information and patterns from scientific literature. It comprises two technologies: text mining and information extraction.

Text Mining – TM, refers to the emerging research area that can be roughly characterized as knowledge discovery in large text collections, thus combining knowledge discovery and text

processing methods. It is concerned mainly with the discovery of interesting patterns such as clusters, associations, deviations, similarities, and differences. *Information Extraction* – IE, aims towards the identification of predefined classes of entities, relations and events that are explicitly mentioned in the literature. [MAT2004], [BEN2005].

The main (among others) research targets of TM and IE from biomedical literature are: (i) identification of genes/proteins, i.e., their names and/or codes, in (prespecified) biomedical text collections, and (ii) induction of potential associations (or, interactions in the case that specific assumptions are made) between them. This is achieved by measuring co-occurrence between protein/gene names in the text references. Respective R&D achievements so-far demonstrate the utility of text-mining approaches – especially with respect to the high accuracy and precision figures exhibited [BUN2005].

17.2 Grid-enabled Data Mining and Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD) is a research field located at the intersection of Machine Learning, Statistics, and Database Theory, and is often defined as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". The characterization of KDD as a process is for example formalized in the CRISP Data Mining Process Model (http://www.crisp-dm.org), which defines the following steps. These steps are usually repeated iteratively, as shown in the following figure:

1. **Business Understanding:** understanding the application domain (e.g. molecular biology and clinical trials). Identifying domain experts, understanding the problem-relevant domain-specific vocabulary, identification of important background knowledge. In particular, understanding the goal of the analysis.
2. **Data Understanding:** understanding the data set that is being examined, i.e. its semantic, variable descriptions, specific data formats. This task is heavily interconnected with business understanding.
3. **Data Preparation:** converting the data from the application-specific format into a format needed for the modelling step, cleaning the data, computation of derived features, feature and subset selection.
4. **Modelling:** the actual analysis of the data using algorithms from Machine Learning or Statistics.
5. **Evaluation:** checking the correctness and applicability of the model in the application context with respect to the goals of the analysis task.
6. **Deployment:** integration of the discovered knowledge in the user domain. Practical application of the model, including pre-processing steps (e.g. advanced data analysis in a clinical trial).

The modelling step has been in the focus of research in Machine Learning and Statistics, where many data analysis algorithms have been developed. Readily available open-source environments like Weka (http://www.cs.waikato.ac.nz/~ml/weka/) or R (http://www.R-project.org) contain a large, steadily growing variety of data mining methods. The other steps in the KDD process are usually treated in a more ad-hoc manner, even though it is widely acknowledged that these steps are very much responsible for the success of Knowledge Discovery projects [Pyle, 1999]. The combination of KDD and Grid technology promises to support these steps and offer an improved development and deployment of data mining solutions.

As the main focus of the KDD aspect of ACGT project will be on Data Mining and Grid technologies, in particular by integrating available open-source data analysis environments, and not on developing learning algorithms per se, the overview in this report will be limited to the aspects of data mining that deal with the process characteristics, in particular distributed data mining and knowledge management for data mining. For an overview on the vast field of data mining as such, we refer to the standard literature, e.g. [HAN2001], [HAS2001], and [MIT1997].

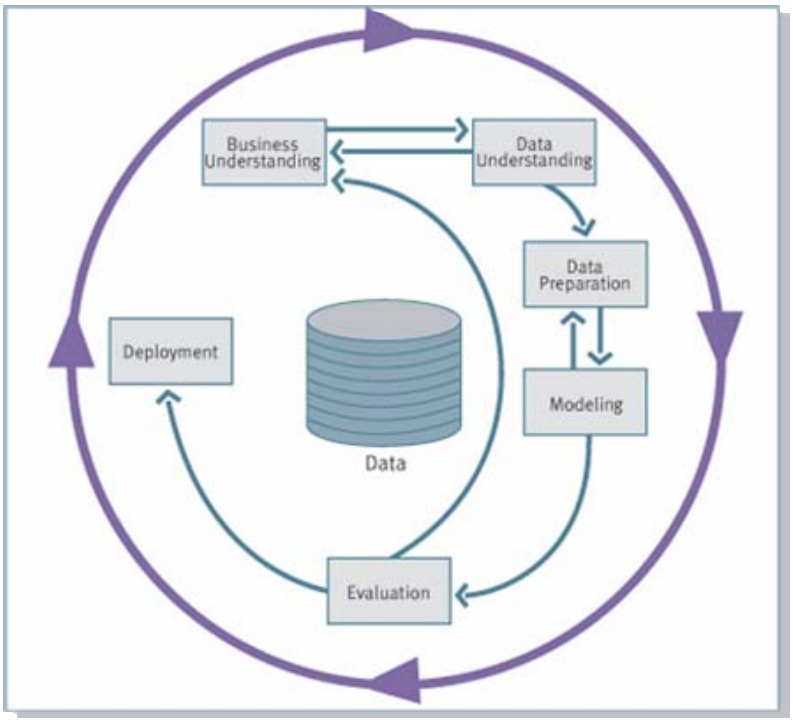


Figure 29: the CRISP Data Mining Process Model

17.3 KDD in Clinical and Genomic Data

Applications of Data Mining and Statistics have intensively been used in clinical and genomic data to deal with the problems of large and complex data, see e.g. [SHO1976], [GUS1997], [LAV1999], [HOR2001], and [MOU2001].

In general, KDD is domain independent in the sense that the actual application is abstracted to tasks like classification, regression, time series analysis; frequent item set mining, or clustering. For example, classification can be applied for automated diagnoses of diseases, regression can be used to predict the optimal dosage of a drug, and time series analysis is an important step in designing online monitoring systems. Frequent item set mining and sequence clustering are both used in bioinformatics to analyze genomic data.

However, each application domain has its specific characteristics that must be taken into account in the decision for or against a specific data mining technique. The task of KDD in clinical and genomic data is characterized by the following aspects:

- High dimensionality: a genomic data set usually consists of a low number of observations (patients), but a high number of measured features for each patient. This makes it hard to identify meaningful correlations and distinguish them from random effects. Usually, a high amount of pre-processing or specialized learning tasks such as frequent item set analysis must be involved to discover meaningful knowledge [BOU2005].

- Heterogeneous data: clinical data sets typically consists of very different types of data, ranging from static information (e.g. age and sex of the patient) over irregularly updated data (e.g. blood screening) up to continuously measured data (online monitoring). Data types can be structured (e.g. blood pressure, heart rate) or unstructured (x-ray images, natural language text). All these types of data may be integrated into a single data set to solve a KDD task.

- Complex background knowledge: doctors use very detailed knowledge about the human organism in their everyday work. Moreover, medical research steadily generates and publishes more knowledge. Identifying and exploiting this vast knowledge is crucial in the analysis of clinical data [MOR2000]. At the same time, data mining techniques can be employed to assist clinicians and researchers in finding and organizing relevant information in the steadily growing stream of medical publications [ANA2004].

17.3.1 Distributed Data Mining

Data Mining is usually a computationally expensive, time-consuming task. Also, often data from various locations has to be integrated into the process. Hence, Data Mining can profit very much from techniques from distributed computing. Although some approaches have been made to parallelize certain data mining algorithms, such as clustering [KAR1999] or association rule mining [ZAK1999], a KDD process is more often characterized by repeatedly applying a single algorithm to multiple data sets, for example in cross-validation experiments, feature selection, and resampling-based approaches, or multiple executions of the same or similar algorithms for parameter tuning and model selection. Hence, most of the KDD processes in practical applications are trivially parallelizable by means of batch processing systems such as Condor [THA2005], or by Remote Method Invocation. Indeed, several extensions for parallelizing jobs in KDD environments exist. These will be discussed in the next section.

An aspect that becomes relevant with respect to distributed data is the distribution of the data preparation step. One can avoid a costly transport of the complete distributed data to a main site, when the data mining step only needs to know certain features of the data. For example, the database-centred Mining Mart system compiles multiple pre-processing steps into a single database query to avoid generating unnecessary intermediate data sets. Limiting the amount extracted data can also circumvent several problems of privacy [VAI2003].

17.3.2 Knowledge Management for Data Mining

Next to the distribution of data and computational resources, one aspect of the Grid is the distribution of knowledge in the sense that it should be easy for each participant to benefit from the diverse services offered from the Grid. From the view of Knowledge Discovery, this amounts to managing knowledge about successful KDD processes and applications [BAR2000].

As a basis for knowledge management for data mining, standards for describing data mining tasks, data and algorithms have been introduced. Common standards greatly simplify the integration, application and maintenance of a system employing a KDD model. The most prominent example is the XML-based Predictive Model Markup Language (PMML)

[RAS2004], which represents and describes data mining models and their associated data types, as well as some pre-processing methods. Another important standard is the Cross-Industry Standard Process for Data Mining [CHA1999], which describes the whole data mining process. Other related standards also exist [GRO2002].

The other side of knowledge management is the ability to use this knowledge. In the context here this describes the ability to execute a process on new data. For example, the Mining Mart project [MOR2004] developed a meta-model (M4) which allows executing a modelled process inside a database by compiling it into an SQL query.

17.3.3 SoA on Grid enabled DM/KDD environments

In the field of data mining we encounter a great variety of applications. While the applications vary a lot regarding the number of data mining algorithms they provide, the algorithms themselves also vary greatly regarding their parameters (number, types, and effects), inputs, outputs and resource requirements. Data mining suites usually represent only a collection of independent data mining algorithms with a unified graphical user interface and underlying data structure. For Grid solutions, this poses the problem of assuring a generic access to a diverse set of algorithms with possibly very different requirements.

Currently, several ongoing projects aiming to couple data mining solutions with Grid computing exist. With four ongoing or already finished projects in this field the data mining toolkit *Weka* (http://www.cs.waikato.ac.nz/ml/weka/) is a very prominent example for the different approaches taken. All projects aim at a tight integration of Grid technology (e.g. WSRF/GT4) into Weka itself rather than developing generic solutions for performing data mining in Grids in general. All projects integrate Grid technology into Weka's GUI itself or require users to start it from the command line rather than providing a high-level elaborated interface to Grid-enabled data mining.

- The **DataMiningGrid project** (Data Mining Tools and Services for Grid Computing Environments, http://www.dataminingGrid.org) is developing tools and services for deploying data mining applications on the Grid. The project is currently implementing a test bed consisting of several prototypical data mining applications developed for diverse sectors such as text mining, gene regulation analysis, and ecological modelling. The DataMiningGrid project seeks new ways to utilize Grid computing technology for applying data mining, especially regarding integration of applications, their execution in the Grid, and user-friendliness.

  The heterogeneity of data mining applications has led the DataMiningGrid approach to avoid implementing specialized services and client-side components for each data mining application which is to be integrated in a Grid environment, if possible. Instead an XML-based meta-data schema has been developed, which allows application developers to describe each single algorithm in detail regarding its options, inputs, outputs, resource requirements and additional information (vendor, version, etc.). This information is stored in a searchable repository, which can be queried at any time to provide an up-to-date view of the applications currently available in the Grid. Furthermore, this information can be used by generic client-side components (e.g. a workflow editor, Web portal) for providing appropriate graphical means for specifying user settings and additional help for specifying input data and resource requirements. The avoidance of implementing specialized services and client-side components for each data mining application also enables sophisticated features such as execution of applications on remote machines, selected automatically by a resource broker, without prior installation of these applications and the ability to utilize thousands of machines for running highly parallel applications.

While the services and client-side libraries developed in this project may also be used to create Web portals, its main focus is on using a workflow editor for specifying a whole set of data mining tasks that are to be performed in the Grid. The generic job template was developed to provide data miners with a view on the whole workflow chain, which resembles those commonly found in data mining suites that feature a workflow editor. It consists of four parts, namely application discovery in the Grid (using the information described above), preparation of input data using OGSA-DAI and/or GridFTP, specification of user settings, and execution of the job. As end-users are expected to be data mining experts rather than Grid-experts, special attention was paid to ensure user-friendliness of the whole system. All components, which require interaction with end-users, provide graphical interfaces, independent whether users are using the workflow editor or other client-side applications. Users are never asked to provide information in any of the many Grid-related formats such as XQuery for application discovery, XPML (used for the job description), SOAP or WSDL. Instead the inputs users provide via the graphical user interface are automatically transformed into the appropriate formats. Additionally, the generic job template hides almost all of the Grid-related details from end-users, allowing them to concentrate on their data mining task at hand rather than having to deal with Grid-related aspects.

- **Discovery Net** [CUR2002] is a project to build a platform for scientific discovery from the data generated by a wide variety of high throughput devices. Discovery Net provides a service-oriented computing model for knowledge discovery, allowing users to connect to and use data analysis software as well as data sources that are made available online by third parties. Discovery Net is based on an open architecture re-using common protocols and common infrastructures such as the Globus Toolkit and the OGSI. It also defines its own protocol for workflows, Discovery Process Markup Language (DPML) which allows the definition of data analysis tasks to be executed on distributed resources.

- **Knowledge Grid** [CAN2003] is a software architecture for parallel and distributed knowledge discovery. It is designed on top of computational Grid mechanisms, provided by Grid environments such as Globus. It includes the basic Grid services such as communication, authentication, information, and resource management. It includes a GUI for developing KD applications.

- **Weka4WS (**http://Grid.deis.unical.it/weka4ws/) is a data mining toolkit developed at the University of Calabria, which uses part of the functionality provided by the Globus Toolkit 4 (GT4). Weka4WS distinguishes between user, computing and storage nodes. While the user nodes provide the standard Weka user interface (Explorer-Panel) into which all Grid functionality is integrated, computing nodes are host WSRF-compliant Web-services, which represent a subset of the data mining algorithms from the original Weka. Weka4WS requires that the respective services are pre-installed on every computing node. As transfer of file based data is realised via GridFTP, the storage nodes must run a GridFTP server. When an analysis is started the respective services deployed on the computing nodes access the files on the storage nodes and read in the required input data.

While this approach is feasible for a quick and tight integration of a single data mining toolkit into a Grid environment, it bears the following crucial disadvantages:

- The middleware employed (GT4) is incompatible to the one widely used in the SIMDAT project (http://www.ecmwf.int/services/grid/simdat/) regarding applied standards (WSRF vs. WS-I), security (proxy certificates vs. no proxy certificates) and

data transport (GridFTP vs. no GridFTP). This probably largely results from the different focuses of both middleware packages. While GT4 focuses mostly on the scientific community, GRIA has been developed from the start to enable inter-enterprise computing such as B2B scenarios.

  - As the name already suggests, Weka4WS is based on WSRF-compliant Web services, meaning that its algorithms are wrapped inside these services. However, as a result these services have to be deployed on all computing nodes prior to starting an analysis and cannot be dynamically transferred and started on a different node. This also means that it is impossible to start an analysis on a cluster controlled by specialized clusterware such as PBS or Condor.

  - As GT4 does not contain any resource scheduler for service invocation and the Weka-services cannot be deployed at runtime on a different machine, compute nodes are prone to "stall" in such cases when many users simultaneously start analyses on the same server.

- **Weka-Parallel (**http://weka-parallel.sourceforge.net/), available at SourceForge.net, does not claim to be Grid-enabled but rather implements distributed cross-validation for all algorithms implemented in Weka relying on Java Remote Method Invocation. Therefore, despite its name, Weka-Parallel is limited to cross-validation only, although significant performance gains have been reported (see:http://weka-parallel.sourceforge.net/report.pdf). Furthermore, it is only applicable when all calculations are done in a multi-processor environment without any firewalls such as a cluster.

- In **Grid-Weka (**http://smi.ucd.ie/~rinat/weka/), which was developed in University College Dublin, execution of Weka tasks can be distributed across several computers in an ad-hoc "Grid". This work is similar to the Weka-Parallel project, but allows for performing more functions in parallel and/or on remote machines. The Weka function is distributed by partitioning the data set, processing several partitions in parallel on different available machines, and merging the results into a single resulting data set. Grid-Weka uses proprietary protocols and relies on a client-server architecture using Java object serialisation for data exchange. Therefore, Grid-Weka is not capable to perform distributed data mining across organizational boundaries, which are usually protected by firewalls. Similar to Weka4WS the application has to be installed on every computer designated to run it. Grid-Weka also provides a mechanism for load balancing. While this is a feasible solution in an environment where no middleware is present, its load balancing would interfere with other schedulers who are often part of standard clusterware, such as Condor. An efficient environment for Grid-Weka would therefore require machines that are exclusively designated to running this package besides those running Condor for general scheduling.

- For the **R** software, several package that offer remote and parallel execution capabilities exists, such as the packages Biopara, Papply, Rmpi, Rpvm, and Taskpr. However, none of these packages provides advanced Grid functionality, but require much work and knowledge on the user side for setting up software configurations, remote hosts and adapting the own code.

17.3.4 Available open source tools

- **R** (http://www.r-project.org) [RDC2005] is an open-source environment for statistical data analysis. It is based on the S language [BEC1988], which is also implemented in

the commercially available Splus system. Next to its base system, several hundred extensions packages for a wide range of applications are available. The R system provides implementations for a broad range of state-of-the-art statistical and graphical techniques, including linear and non-linear modelling, cluster analysis, prediction, hypothesis tests, resampling, survival analysis, and time-series analysis. R code can either be implemented in the R language, or using R's interface to C libraries.

- **Bioconductor** (http://www.bioconductor.org) [GEN2005] is an open source and open development software project for the analysis and comprehension of genomic data. The project was started in the fall of 2001. Bioconductor is primarily based on the R programming language. The Bioconductor project aims to provide access to a wide range of powerful statistical and graphical methods for the analysis of genomic data. Analysis packages are available for: pre-processing Affymetrix and cDNA array data; identifying differentially expressed genes; graph theoretical analyses; plotting genomic data. In addition, the R package system itself provides implementations for a broad range of state-of-the-art statistical and graphical techniques, including linear and non-linear modelling, cluster analysis, prediction, resampling, survival analysis, and time-series analysis. In addition to the main software there are a large number of meta-data packages available. They are mainly, but not solely oriented towards different types of microarrays. The Bioconductor project also provides software for associating microarray and other genomic data in to biological metadata from web databases such as GenBank, LocusLink and PubMed.

- **Weka** (http://www.cs.waikato.ac.nz/~ml/weka/) [WIT2005] from the University of Waikato is a popular machine learning environment implemented in Java. The algorithms can either be applied directly to a dataset from within a graphical user interface, or called from other Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

- **Yale** (http://yale.cs.uni-dortmund.de) [RIT2001], which was developed at the University of Dortmund, Germany, YALE is an environment for machine learning experiments and data mining which is similar to Weka. A modular operator concept allows constructing experiments can by combining a large number of arbitrarily nestable operators. In particular, operators from Weka can be directly imported into the Yale environment. YALE provides more than 350 operators including over 100 learning schemes. Yale contains a plug-in mechanism for easily integrating new Java classes into the core system. Experiments can be developed, executed and visualized in a GUI, but can also be saved in XML format and executed by a command line version.

17.3.5 Generation of indicative DM/KDD scenarios for the ACGT clinico-genomic trials

The development of detailed KDD scenarios must of course be closely coordinated with the goals of the clinico-genomic trials as defined by clinicians. In general, three different user groups of the ACGT platform can be distinguished, each of which with a different interest and expertise in the analysis of clinico-genomic data. This gives rise to the following three scenarios.

- **Clinician's scenario**: the clinician is assumed to have very detailed knowledge of the goal of the study and the underlying medical and biological processes, but little expertise in statistical analysis and data mining. Accordingly, in this scenario the user is assumed to define the type of analysis task, e.g. the prediction of a certain variable,

plus all relevant data. The task of the system will be to automatically choose an optimal workflow and modelling approach from a set of pre-defined templates. For example, for a classification task the user may describe his data set, in particular the target variable, define additional databases that contain important information, and select a workflow template from a set of alternatives. The task of the system will then be to automatically evaluate several applicable classification algorithms, tune their parameters, and perform automatable pre-processing steps like feature selection in order to optimize the predictive performance. As this scenario is computationally very expensive, the system should automatically make use of the available Grid resources.

- **Molecular Biologist's / Bioinformatician's scenario**: in this scenario the users are supposed to have more detailed knowledge on the application of data mining and statistical methods, but without detailed knowledge of Grid internals or on the development of novel data mining algorithms. Hence, in this scenario it is assumed that the users interests lies in developing new workflows. The system should provide the workflow developer with the capabilities to search for available algorithms, data sources, and other resources provided by the Grid. Further, the workflow designer should be able to publish his workflow, such that it can be used in a clinician's scenario. The workflow designer may also want to find workflows similar to his own, and compare the results of his workflow on a set of available data sets / analysis tasks.

- **Data Miner's scenario**: in a clinico-genomic trial it may turn out that available data mining algorithms are not sufficient to achieve the goal of the trial. Then it will be necessary to develop and implement novel data mining approaches. In this scenario the user is assumed to be a data mining specialist with limited knowledge about the application domain of clinico-genomic trials. The data miner should be able to construct new algorithms using pre-defined functions that allow him to easily make use of the available Grid resources, e.g. for the development of parallel algorithms. Additionally, he may need to exploit domain knowledge that is available in the form of ontologies to guide the development of sensible new analysis methods.

17.4 2D Visualisation Tools

This section gives an overview of a range of tools that are used in 2D visualization. It is mainly focused on tools that could potentially be used in the context of bibliographic corpora analysis by visual means.

Traditionally bibliographic analysis was carried out through the use of Boolean based keywords searches. In Biovista's BioLab Experiment Assistant (BEA) [WEB2] software, a new paradigm is introduced in which a user can visually analyze a corpus dealing with entities. In such an environment proper visualization techniques are of paramount importance since they allow a user to accurately and productively acquire its target set of articles from massive corpora.

A widely used approach in visualizing concepts contained within atomic items in corpora is the use of weighted graphs. The weights in most cases are the outcome of statistical analysis of affinity within concepts, such as similarity indices etc.

Within this scope the remainder of the document outlines technical capabilities of major visualization libraries that could be used for the above named purpose.

17.4.1 yFILES

Overview

yFiles is a Java class library that provides algorithms and components enabling the analysis, visualization, and the automatic layout of graphs, diagrams, and networks [WEB1]. yFiles is a commercial product.

Component structure

The yFiles libraries are divided in three main parts which while they remain independent (feature a clear layer of functional separation) they can be interchangeably combined to provide the user with a complete class library for both the representation and visualization of graphs and networks. Its main components are:

- **yFiles Basic:** yFiles Basic contains essential classes and data types for graph analysis tasks. It ensures the highly efficient implementation of advanced data types such as graph and priority queue. Furthermore, it makes available a wide variety of graph and network algorithms which in turn form an indispensable toolkit for a range of network analysis tasks.

- **yFiles Viewer:** yFiles Viewer builds upon the Basic package. It offers you a powerful graph viewer component, which is showcased in the yEd Graph Editor application and other Swing-based GUI elements. Other impressive features of the Viewer package are its ability to support diverse graph formats (GML, YGF, JPG, GIF) and its excellent printing capabilities.

- **yFiles Layout:** yFiles Layout also builds upon the Basic package. It offers a perfect suite of graph layout algorithms, which deliver you unrivaled opportunities. Diverse layout styles including hierarchic, orthogonal, or circular are provided as easy to integrate components that can be configured programmatically to suit most layout demands. In addition, yFiles provides edge routing algorithms that make it possible to easily route edges into existing diagrams. The different layout styles also now include several incremental algorithms, for example incremental hierarchical layout.

Main Application Domains

yFiles provides a user with the essential building blocks for Java applications that are needed to analyze, visualize, edit, or automatically draw graphs, diagrams, or networks. According to the manufacturer, application domains in which yFiles has been used include

- biochemical network analysis and visualization
- business process modelling
- data mining
- database management and modelling
- network management
- social networks
- software engineering
- workflow management
- WWW visualization

- visual programming

Example Output

The following is a sample of output capabilities of yfiles.

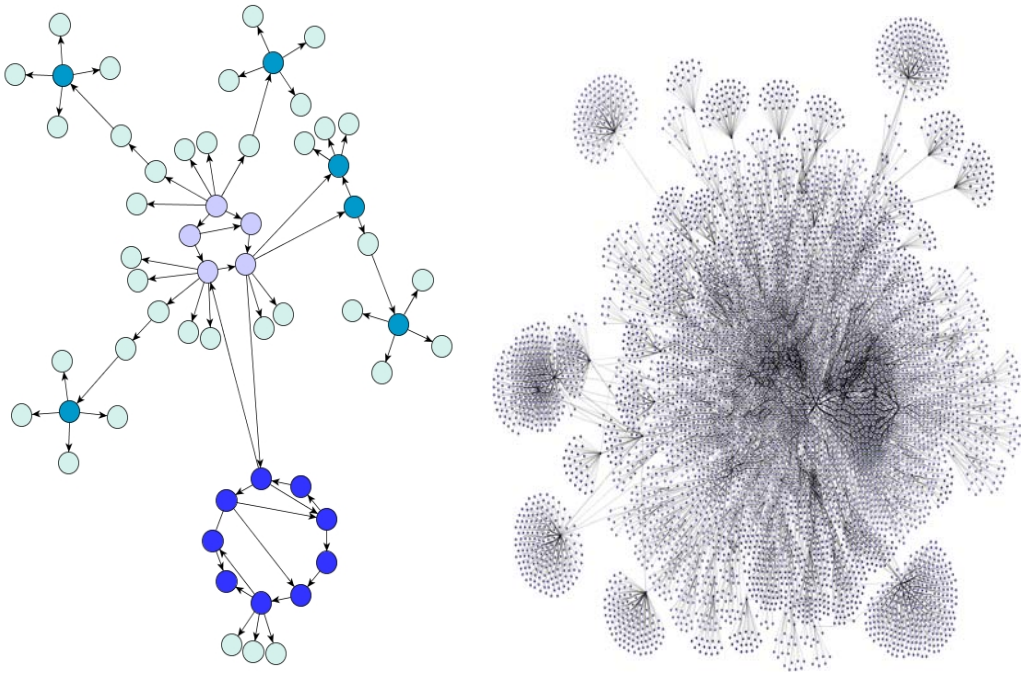


Figure 30: yFiles Examples

Large graphs (containing tens of thousands of nodes) are possible. These can be annotated in a variety of ways and at the same time a variety of layout algorithms are available to the programmer.

Layout quality depends on the amount of time allocated programmatically to yfiles for the calculation of the optimal positioning and minimal overlaps.

17.4.2 Tom Sawyer

Overview

The Tom Sawyer (TS) product family enables the user to develop graph and network analysis applications. As with yFiles it provides a variety of tools required for the representation and visualization of complex graphical information. The Tom Sawyer range of tools, provide quite sophisticated clustering, graph traversal, path analysis, dependency analysis, impact analysis, network analysis and other high-value functions that can help “improve analytic decision making”.

Component Structure

The TS libraries are, in similar style to yFiles, divided in three main logical parts. These can be combined to provide the user with a complete set of tools for both the representation and visualization of graphs and networks. Its main components are:

- **Tom Sawyer Analysis:** This class library provides the user with the ability to represent graphs and networks in a non visual manner. They are the cornerstone

structures used in graph representation as well as defining fundamental operation on such sets.

- **Tom Sawyer Visualization:** The Tom Sawyer Visualization class library enables you to develop graph visualization applications quickly and efficiently. It provides the means for creating applications with sophisticated graph display, viewing and editing technologies presented in an eye-catching, intuitive graphical user interface. You can customize both the display and the interactive behaviours of an application using industry standard components, such as toolbars, menus and mouse-event handling.

- **Tom Sawyer Layout:** The Tom Sawyer Layout component adds scalable graph layout capabilities to your applications. Graph layout technology reveals complex relationships in data by automatically computing diagrams. These diagrams expose the underlying graph structures as well-organized drawings that you can immediately understand. And because the technology is portable and flexible, you can easily integrate it with your own database, display and graphics software.

Main Application Domains

According to the manufacturer, main application domains of the product include:

- biochemical network analysis and visualization
- process modelling
- development and representation of social networks
- software engineering
- workflow management
- visual programming
- bibliographic Analysis

17.4.3 JGraph

Overview

The core JGraph library is a feature-complete Java Graph Visualization component designed for maximum flexibility and small jar size. 100% pure Java, JGraph has been used in thousands of applications world-wide, both rich client and server-side.

JGraph is a feature-rich open source graph visualization library written in Java. JGraph is written to be a fully Swing compatible component, both visually and in its design architecture. JGraph provides a range of graph drawing functionality for client-side or server-side applications. JGraph has a simple, yet powerful API enabling you to visualize, interact with, automatically layout and perform analysis of graphs.

Component Structure

JGraph follows a single component architecture at a conceptual level but it logically divided in several packages.

The main package is JGraph itself which comprises the basic JGraph swing component. The following packages make up the rest of its functionality:

Java Package Name	Functionality
org.jgraph	**Basic JGraph class**
org.jgraph.event	**Graph Event Models**
org.jgraph.graph	**Graph Structure and nodes**
org.jgraph.plaf	**Graph UI delegate component**
org.jgraph.util	**General purpose utilities**

Main Application Domains

According to the creators, the main application domains are:

- Process diagrams, workflow and BPM visualization, flowcharts, even traffic.
- Database and WWW visualization, networks and telecoms displays, mapping applications and GIS.
- UML diagrams, electronic circuits, VLSI, CAD, social networks and data mining.
- Biochemistry, ecological cycles, entity and cause-effect relationships and organizational charts.

Example Output

The following are screenshots of the JGraph tool.

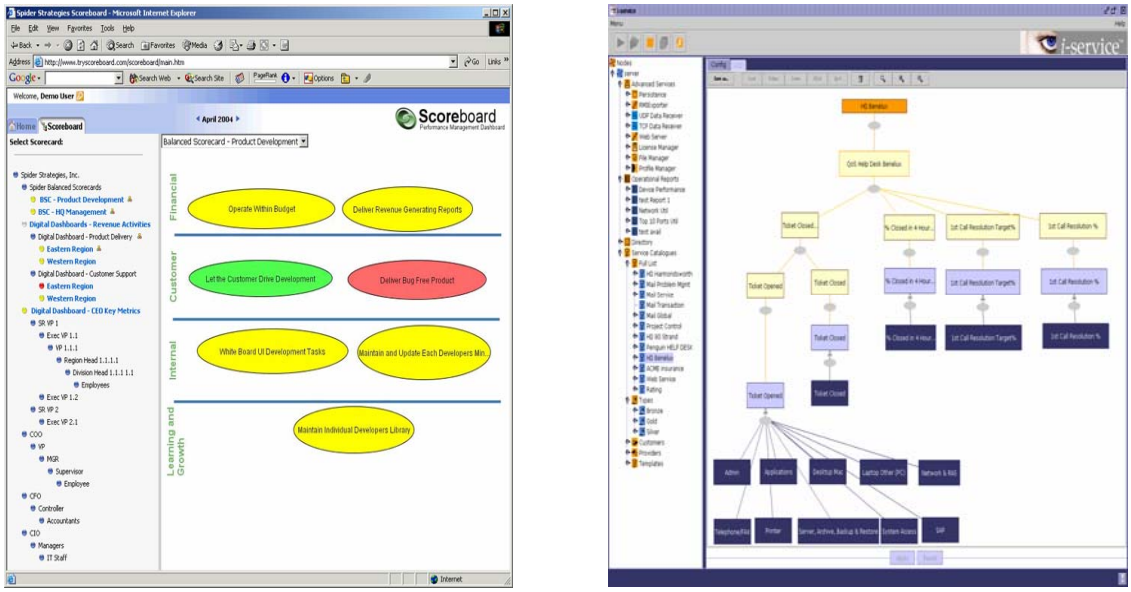


Figure 31: JGraph Examples

17.4.4 JViews

Overview

JViews from ILOG [WEB4] is a range of tools that allow a user to visually represent several and in many cases quite different domains. These vary from map viewing to organizational charts. It is not limited in 2D visualization but has 3D support as well.

Component Structure.

JViews is comprised of the following 4 libraries:

- **ILOG JViews Diagrammer**

JViews Diagrammer is a comprehensive set of Java components, tools and libraries for creating diagram-based editing, visualization, supervision and monitoring tools. JViews Diagrammer offers variety of tools for Java developers. Examples include:

1. A high-level model-based diagram component that includes diagram supervision and modelling capabilities
2. A point-and-click editor, Diagrammer Designer that accelerates diagram configuration through visual tools.
3. Pre-built diagram editors, including a business process modelling notation (BPMN) editor and a drawing editor, delivered with source code.
4. A complete software development kit (SDK) that allows the user to extend the diagram component.

The extensive graph layout capabilities of JViews Diagrammer empower allow the user to create a fully customized diagram layout. Features include a series of advanced graph, link, and label layout algorithms that stay clean and readable - even at design time.

- **JViews Maps**

ILOG JViews Maps provides the tools for adding essential interfaces to Web-based displays. The user can quickly deploy accurate, functional maps that power critical Java monitoring applications. It includes a complete API for map readers; projections; load-on-demand; and a MapBuilder tool for configuring maps. Custom graphic objects represent underlying business objects. Objects are shown in correct geographic positions. An SDK lets the user to manage maps in Java, and display georeferenced graphics objects atop maps.

- According to the manufacturer, its main features include:
- Georeferencing for easy placement of assets in proper locations
- Multiple projections that support the most popular methods of projecting the earth's surface
- Mix-and-match vector and raster data in the same map
- Native support for common map formats
- Load-on-demand for efficiently handling very large sets of map data
- Connection to the popular Oracle Spatial Map server (http://www.oracle.com/technology/products/spatial/index.html)
- MapBuilder, an editor for defining map data used in applications

- Complete customization of all map features and all objects atop a map
- Multiple Web-deployment options, including Java clients and thin DHTML clients
- Map readers for standard and commercial data formats including Oracle Spatial, ShapeFile, MID/MIF, DTED, CADRG, GeoTIFF, GIF and JPEG

- **JViews Gantt**

JViews Gantt is a complete set of Java components, tools and APIs for viewing and editing schedules. JViews Gantt also includes JViews Charts, allowing the user to display any type of chart.

JViews Gantt includes resource and task-oriented views, as well as resource load charts. Comprehensive editing and viewing features provide complete display control. A load-on-demand mechanism accommodates very large data sets. Printing options expedite hard copies.

Its Gantt's chart Designer tool lets the user personalize Gantt charts using a point-and-click editor. A comprehensive SDK, with powerful APIs, is also included. It implements a Swing-like model view controller (MVC) architecture, providing clear separation between data and its screen representations.

The data model is completely open and extensible, allowing it to connect with any other application component. Notifications are automatic and transparent - when the data model changes, views are updated; when the user interacts with views, the model is changed.

An XML extension called Schedule Data eXchange Language (SDXL) can be used to serialize Gantt and schedule charts -- this is particularly useful for displaying and interacting with Gantt charts while disconnected from the scheduling application. Users can transfer an SDXL document, work offline, and then upload the document to the application when reconnected. It can also exchange scheduling data with other programs -- XSLT can be used to translate XML-based scheduling languages to SDXL and vise-versa.

- **ILOG JViews Charts**

JViews Charts provides complete support for Web-based charting including the incorporating of charts into Web GUIs. JViews Charts offerss dynamic, interactive charts to Web GUIs without compromising usability or performance. It is highly customizable, a full complement of custom parameters makes charts look and behave as users want. Parameters include:

- Linear or logarithmic scales
- Multiple axes, flipped axes, swapped axes
- Smart tick mark labels that don't overlap
- User definable Grids and scales.
- Labels on data points, user-definable data point markers
- Legends to describe charts
- Behaviours for pan, zoom, edit, pick and more
- Render the data in 2D or 3D
- "Video" zooming, "magnifier lens" zooming, "tool tips" and more
- Even end users can customize their displays

JViews Charts supports DHTML clients as well as traditional Java clients. Applications can drive both types of displays -- using the same data model to drive both displays at the same time. All charts are available as Java Server Faces (JSF) that make interactive Web charts easy to build. These charts, available as both DHTML and SVG clients, leverage AJAX concepts for greater interactivity and reduced roundtrips to the server.

17.4.5 Discussion

While all the products featured above are of an extremely high standard fro the purpose of visual bibliographic analysis, it is our view through empirical evaluation that yFiles has a cutting edge over its competitors. This is mainly so due to the following:

- Small class library footprint. This makes light application client smaller in size.
- Modular design of its architecture. All of its component can be used with either custom structures or combined with the ones provided by the class libraries.
- Ability to deal with very large graphs in an efficient manner.
- Customizable quality of the results of its layouters depended on time available.
- Direct output support for GML as well as the integration and serialization of custom structures as part of the graphs.
- Straight forward design of class libraries.

Tom Sawyer follows a close second meeting most of the above advantages of yFiles but does so in a slightly more complicated way. It is also not as economically efficiently as yFiles for the provision of the same level of Functionality.

JGraph lacks several major capabilities especially in terms of lay outing as well as exporting capabilities. It also does not support nesting of nodes. It does though present the advantage of the smaller footprint of the featured components and is an open source tool.

Finally JViews provides tools to cover a very wide range of applications but most of its components lack the depth of its competitors which remain highly focused to just the area of graph and network representation and visualization.

17.5 References

[ANA2004] Ananiadou S., Friedmann, C. and Tsujii, J. (2004), “Named Entity Recognition in Biomedicine”, *Journal of Biomedical Informatics,* 37(6), 393-528.

[BAR2000] Bartlmae, K. and Riemenschneider, M. (2000), “Case Based Reasoning for Knowledge Management in KDD Projects”, *Proc. of the Third Int. Conf. of Practical Aspects of Knowledge Management.*

[BEC1988] Becker, R., Chambers, J., and Wilks, A. (1988), “The New S Language”, Chapman & Hall, London.

[BOU2005] Boulicaut, Jean-Francois (2005), “Towards Inductive Databases for Gene Expression Data Analysis”, in: *Local Pattern Detection,* Morik, Boulicaut, Siebes (eds.), Springer.

[CAN2003] Cannataro, M., Talia, D. (2003), "KNOWLEDGE GRID: An Architecture for Distributed Knowledge Discovery", *Communications of the ACM*, January 2003.

[CHA1999] Chapman, Pete, Clinton, Julian, Khabaza, Thomas, Reinartz, Thomas, and Wirth, Rüdiger (1999), “The CRISP-DM Process Model”.

[CUR2002] Curcin, V., Ghanem, Y., Guo, M., Kohler, M. Rowe, A., Syed, J., Wendel, P. (2002), “Discovery Net: Towards a Grid of Knowledge Discovery”, in: *KDD-*

2002. The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[GEN2005] Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S. (2005), "Bioinformatics and Computational Biology Solutions Using R and Bioconductor", Springer.

[GRO2002] Grossmann, R. Hornick, M., and Meyer, G. (2002), "Data Mining Standards Initiatives", *Communications of the ACM*, 45(8).

[GUS1997] Gusfield, Dan (1997), "Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biolog", University of Cambridge.

[HAN2001] Hand, David, Mannila, Heikki and Smyth, Padhraic (2001), "Principles of Data Mining", MIT Press.

[HAS2001] Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome (2001), "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Springer.

[HOR2001] Horn, Werner (2001), "AI in Medicine on its way from knowledge-intensive to data-intensive systems", *Artificial Intelligence in Medicine,* 23(1), 5-12.

[KAR1999] Kargupta, Hillol, Huang, Weiyun, Sivakumar, Krishnamoorthy and Johnson, Erik (1999), "Distributed Clustering Using Collective Principal Component Analysis", *Knowledge and Information Systems.*

[LAV1999] Lavrac, Nada (1999), "Selected Techniques for data mining in medicine", *Artificial Intelligence in Medicine,* 16(1), 3-23.

[MIT1997] Mitchell, Tom (1997), "Machine Learning", McGraw Hill.

[MOR2000] Morik, Katharina, Imhoff, Michael, Brockhausen, Peter, Joachims, Thorsten and Gather, Ursula (2000), "Knowledge Discovery and Knowledge Validation in Intensive Care", *Artificial Intelligence in Medicine,* 19(3), 225-249.

[MOR2004] Morik, K. And Scholz, M. (2004), "The MiningMart Approach to Knowledge Discovery in Databases", in: *Intelligent Technologies for Information Analysis,*Zhong and Liu (eds.), 47-65.

[MOU2001] Mount, David W. (2001), "Bioinfomatics: Sequence and Genome Analysis", Cold Spring Harbor Laboratory Press.

[PYL1999] Pyle, Dorian (1999), "Data Preparation for Data Mining", Morgan Kaufmann Publishers.

[RAS2004] Raspl, Stefan (2004), "PMML Version 3.0---Overview and Status", *Proc. of the Workshop on Data Mining Standards, Services and Platforms at the 10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining,* 18-22.

[RDC2005] R Development Core Team (2005), "R: A language and environment for statistical computing", R Foundation for Statistical Computing.

[RIT2001] Ritthoff, O., Klinkenberg, R., Fischer, S., Mierswa, I., Felske, S. (2000), "Yale: Yet Another Machine Learning Environment", in Proc. *LLWA 01*, 84-92.

[THA2005] Thain, D., Tannenbaum, T., and Livny, M. (2005), "Distributed Computing in Practice: The Condor Experience", *Concurrency and Computation: Practice and Experience*, Vol. 17, No. 2-4, 323-356.

[VAI2003] Vaidya, Jaideep and Clifton, Chris (2003), "Privacy-Preserving K-Means Clustering over Vertically Partitioned Data", *Proc. Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

[WEB1] http://www.yworks.com/

[WEB2] http://www.biovista.com/pub/bea

[WEB3] http://www.tomsawyer.com/

[WEB4] http://www.ilog.com/

[WEB5] http://www.jgraph.com/

[WIT2005] Witten, Ian and Frank, Eibe (2005), "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco.

[ZAK1999] Zaki, Mohammed (1999), "Parallel and Distributed Association Mining: A Survey", *IEEE Concurrency,* 7(4), 14-25.

18 Metadata & Metadata standards

18.1 Introduction

ACGT focuses on semantic integration of data but also on the discovery, integration, and management of searable data assets (i.e. data and toolsl operating on such data). As a result the issue of metadata becomes of paramount importance for the succesfull achievement of the project objectives. This Chapter reviews relevant state of the art and provides information on the most important and relevant to ACGT international initiatives.

Metadata is often called ‘data about data’. More precisely, it is the underlying definition or structured description of the content, quality, condition or other characteristics of data. The term was introduced intuitively, i.e. without exact definition. Because of that today there is a whole variety of definitions. The most common one is the literal translation:

- Metadata is data on data.

As for most people the difference between data and information is merely a philosophical one of no relevance in practical use. Existing definitions include:

- "Metadata is information about data"
- "Metadata is information about information".

There are also more sophisticated definitions such as:

- "Metadata is structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities." [DUR1985] and
- "[Metadata is a set of] optional structured descriptions that are publicly available to explicitly assist in locating objects." [KIM1998].

18.2 Semantic Web

A lot of efforts in metadata descriptions are centered on the semantic web and its technologies, namely RDF and OWL. RDF is a simple data model for referring to objects ("resources") and how they are related. An RDF-based model can be represented in XML syntax. RDF Schema is a vocabulary for describing properties and classes of RDF resources, with a semantics for generalization-hierarchies of such properties and classes. OWL adds more vocabulary for describing properties and classes: among others, relations between classes (e.g. disjointness), cardinality (e.g. "exactly one"), equality, richer typing of properties, characteristics of properties (e.g. symmetry), and enumerated classes.

In the area of the Semantic Web Services the following technologies and standards are relevant to the semantic description of web services:

- **UDDI** (Universal Description, Discovery and Integration - http://www.oasis-open.org/committees/uddi-spec/doc/tcspecs.htm) allows the discovery of potential

business partners on the basis of the services they provide. Each business description in UDDI consists of a businessEntity element that describes a business by name, a key value, categorization, services offered (businessServices) and contact information for the business. Each businessService element contains descriptive information such as names and descriptions, and also classification information describing the purpose of the relevant Web service. Using UDDI, a Web service provider registers its advertisements along with keywords for categorization. A Web services user retrieves advertisements out of the registry based on keyword search. So far, the UDDI search mechanism relied on predefined categorization through keywords, but more recently specifications to use OWL in UDDI are emerging as a uniform way to express taxonomies business taxonomies.

- The **Dublin Core** metadata element set (http://dublincore.org/documents/dcmi-terms/) is a standard (NISO Standard Z39.85-2001 and ISO standard 15836:2003) that provides a simple and standardized set of conventions for describing networked "resources" through a set of elements such as 'Creator', 'Subject', 'Title', 'Description', and so on. Dublin Core is widely used to describe digital materials such as video, sound, image, text and composite media like web pages. Implementations of Dublin Core are typically XML and RDF based.

- **WSDL-S** (http://www.w3.org/Submission/WSDL-S/) is a means to add semantics inline to WSDL. It is actually an extension to WSDL 2.0 but can be used for WSDL 1.1. According to WSDL-S inputs and outputs of WSDL operations are annotated with domain concepts while the operations themselves are annotated with preconditions and effects (post conditions). Also the service's interface is annotated with category information which could be used while publishing services in registries such as UDDI. The semantic domain model used is external to these annotations and could be expressed in OWL or other ontology language of choice.

- **OWL-S** (http://www.daml.org/services/owl-s/), formerly DAML-S, builds on top of OWL and allows for the description of a Web service in terms of a Profile, which tells "what the service does/provides", a Process Model, which tells "how the service works", and a Grounding, which tells "how to access the service" [MAR2004]. The service profile describes what is accomplished by the service, any limitations on service applicability and quality of service, and requirements that the service requester must satisfy in order to use the service successfully. The process model gives details about the semantic content of requests, the conditions under which particular outcomes will occur, and, where necessary, the step by step processes leading to those outcomes. In the process model a service can be described as an atomic process that can be executed in a single step or a composite process that, similar to a workflow, can be decomposed in other processes based on control structures like 'if-then-else' and 'repeat-while'. Finally, Grounding descriptions supply information about the communication protocol and other transport information (such as port numbers) and the message formats and serialization methods used in contacting the service. The only currently specified grounding mechanism is based on WSDL 1.1 and will be extended to WSDL 2.0 as soon as it's finalized.

- The **Semantic Web Services Framework** (SWSF - http://www.daml.org/services/swsf/), initiated by the Semantic Web Services Initiative (http://www.swsi.org/), includes the Semantic Web Services Language (SWSL) and the Semantic Web Services Ontology (SWSO). SWSL is a logic-based language for specifying formal characterizations of Web service concepts

and descriptions of individual services. SWSO is an ontology of service concepts defined using SWSL and incorporates a formal characterization ("axiomatization") of these concepts in first-order logic.

- **WSMO** (Web Services Modeling Ontology - http://www.wsmo.org/index.html) defines the modeling elements for describing several aspects of Semantic Web services. These elements are Ontologies, which provide the formal semantics to the information used by all other elements, Goals which specify objectives that a client might have when consulting a Web service, Web services that represent the functional and behavioral aspects which must be semantically described in order to allow semi-automated use, and Mediators that are used as connectors and they provide interoperability facilities among the other elements. It also defines the Web Service Modelling Language (WSML) which formalizes WSMO and aims to provide a rule-based language for the Semantic Web.

- **BioMOBY (**http://www.biomoby.org/) is a Web Service interoperability initiative in the field of bioinformatics aiming to facilitate the integration of web-based bioinformatics resources. Currently there are two approaches to achieve such integration: The first approach, based on the Web Services paradigm, is referred to as "MOBY Services" (MOBY-S), while the second one is called "Semantic MOBY" (S-MOBY - http://www.semanticmoby.org/) and is based on concepts from the Semantic Web. MOBY-S uses a set of simple, end-user-extensible ontologies as its framework to describe data semantics, data structure, and classes of bioinformatics services. These ontologies are shared through a Web Service registry system, MOBY Central, which uses the ontologies to semantically bind incoming service requests to service providers capable of executing them. S-MOBY on the other hand employs RDF and OWL and the document oriented infrastructure of the WWW (the GET/POST methods of HTTP) for publishing and retrieving information from its discovery servers.

18.3 Metatada and Semantic Grid

Semantic Grid [GOB2004] is the application of the principles of the Semantic Web to the Grid environment so that "information and services are given well-defined meaning, better enabling computers and people to work in cooperation". For the realization of the Semantic Grid the technologies of Semantic Web, especially RDF and RDF-Schema, are reused as they seem to be a very natural framework for representing information about Grid resources. In this area of Grids and Semantic Grids we have found interesting the following efforts:

- The Metadata Catalog (**MCAT-** http://www.sdsc.edu/srb/index.php/MCAT) is an information catalog system implemented at San Diego Super Computing Center as an important component of the Storage Resource Broker (SRB) system. The MCAT catalog provides an abstraction mechanism so that users can access 'data objects' (e.g. files) via attributes or logical name, without knowing the physical name or location of a data object. The metadata managed by MCAT include a core-level of meta information, such as location, size, creation date, etc, and also domain-dependent meta information. The architecture for enriching MCAT to support semantic search and discovery is described in [JEF2006].

- The Monitoring and Discovery System (**MDS**) in Globus Toolkit (http://www.globus.org/toolkit/mds/) is a suite of web services to monitor and discover resources and services on Grids. MDS services provide query and

subscription interfaces to arbitrarily detailed resource data and a trigger interface that can be configured to take action when pre-configured trouble conditions are met. The service information is published in MDS as WSRF resource properties and potentially could contain metadata about the service. The MDS though is focused more on the mechanism to disseminate and gather information on Grids rather than the information model to describe services or resources. S-MDS [PAH2006] is a recent proposed enhancement to MDS for supporting the semantic metadata of WSRF Resources, it's based on OWL-S and the information is stored in RDF format.

- The **OGSA-DAI-RDF** middleware (http://www.gtrc.aist.go.jp/dbGrid/index.html) by the National Institute of Advanced Science and Technology (AIST) of Japan extends OGSA-DAI access to RDF database systems. The query language interface is based on SPARQL query language.

- **Semantic OGSA** (S-OGSA) [ALP2006] has been proposed as a light weight extension to OGSA to support Semantic Grid Services and will be the reference architecture for the semantic Grid in the OntoGrid (http://www.ontoGrid.org/) project.

- Also from the OntoGrid project, the use of technologies coming from the Peer-to-Peer (P2P) networks is investigated. In [KOU2006] it is proposed to view resource discovery in Semantic Grids as distributed RDF query answering using Distributed Hash Tables (DHT). In this distributed resource discovery data and metadata are described using RDF while RQL (http://www2002.org/CDROM/refereed/329/) is used as the query language and RUL (http://www.intelligence.tuc.gr/publications/rul.pdf) is used for updating the RDF information.

- **OWL-WS** ("OWL for Workflows and Services") [BEC2005] is a workflow and service ontology supported by the NextGrid project. It is based on OWL-S with various extensions to provide abstract definition of workflows so that a task's bindings (endpoints, and so on) can be specified at run time, to support the substitution of an abstract workflow with a concrete one at run time, and to support higher order workflows so that a task can have a workflow as input and return another workflow as output.

18.4 The ISO/IEC 11179 standard

For ISO/IEC 11179, **metadata** is defined to be data that defines and describes other data. This means that metadata are data, and data become metadata when they are used in this way. This happens under particular circumstances, for particular purposes, and with certain perspectives, as no data are always metadata. The set of circumstances, purposes, or perspectives for which some data are used as metadata is called the **context**. So, metadata are data about data in some context.

A conceptual model of an MDR for describing data is provided in ISO/IEC 11179-3. The requirements and procedures for the ISO/IEC 11179 aspects of registration are described in ISO/IEC 11179-6. For actual metadata registries, there may be additional requirements and procedures for registration, which are outside the scope of ISO/IEC 11179. Rules and guidelines for providing good definitions and developing naming conventions are described in

ISO/IEC 11179- 4 and ISO/IEC 11179-5, respectively. The role of classification is described in ISO/IEC 11179-2.

Recommendations and practices for registering data elements are described in ISO/IEC TR 20943-1. Recommendations and practices for registering value domains are described in ISO/IEC TR 20943-3. An MDR contains metadata describing data constructs. The attributes for describing a particular data construct (e.g., data elements) are known, collectively, as a metadata object. When the attributes are instantiated with the description of a particular data construct, they are known as a metadata item.

Registering the metadata item (i.e., entering the metadata into the MDR) makes it a registry item. If the registry item is also subject to administration (as in the case of a data element), it is called an administered item.

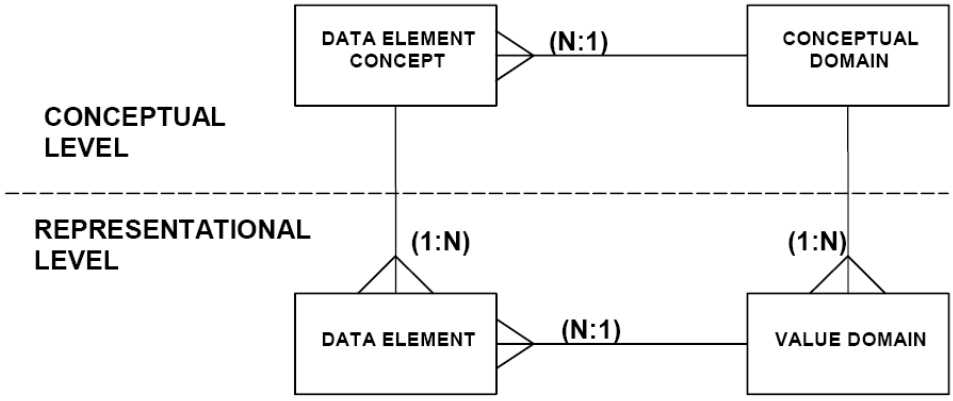


Figure 32: Overview Model for ISO/IEC 11179 Metadata Registry

The model presented and described in ISO/IEC 11179-3 (*Registry metamodel and basic attributes*) is general. It is a representation of the human understanding of the metadata needed to describe **data constructs**, including the relationships that exist among that metadata, and not necessarily how the metadata will be represented in an application of an MDR. A model of this kind is called a **conceptual model**. Conceptual models are meant for people to read and understand.

Models that describe metadata are often referred to as **metamodels**. The conceptual model presented in ISO/IEC 11179-3 is a metamodel in this sense.

18.4.1 Basic principles for applying ISO/IEC 11179

Each part of ISO/IEC 11179 assists in a different aspect of metadata creation, organization, and registration; and each part shall be used in conjunction with the other parts. ISO/IEC 11179-1 establishes the relationships among the parts and gives guidance on their usage as a whole. ISO/IEC 11179-3 specifies metadata items a registration applicant shall provide for each object to be registered. Detailed characteristics of each basic attribute are given. Because of their importance in the administration of metadata describing data constructs, three of the attributes (name, definition, and identification) are given special and extensive treatment in two documents. ISO/IEC 11179-4 shall be followed when constructing data definitions. Identification and naming shall follow principles set forth in ISO/IEC 11179-5. ISO/IEC 11179-2 specifies a set of attributes for use in the registration and administration of classification schemes and their components. Metadata items are registered as registry

items and administered as administered items in an MDR. ISO/IEC 11179-6 provides guidance on these procedures.

18.4.2 Fundamental model of data elements

For the purposes of ISO/IEC 11179, a **data element** is composed of two parts:

- **Data element concept** – A DEC is **concept** that can be represented in the form of a **data element**, described independently of any particular representation.
- **Representation** – The representation is composed of a value domain, datatype, units of measure (if necessary), and representation class (optionally).

From a data modelling perspective and for the purposes of ISO/IEC 11179, a data element concept may be composed of two parts:

- The **object class** is a set of ideas, abstractions, or things in the real world that can be identified with explicit boundaries and meaning and whose properties and behaviour follow the same rules
- The **property** is a characteristic common to all members of an object class.

Object classes are the things for which we wish to collect and store data. They are concepts, and they correspond to the notions embodied in classes in object-oriented models and entities in entity-relationship models. Examples are cars, persons, households, employees, and orders. Properties are what humans use to distinguish or describe objects. They are characteristics, not necessarily essential ones, of the object class and form its intension. They are also concepts, and they correspond to the notions embodied in attributes (without associated datatypes) in object-oriented or entity-relationship models. Examples of properties are colour, model, sex, age, address, or price.

An object class may be a **general concept**. This happens when the set of objects corresponding to the object class has two or more members. The examples in the previous paragraph are of this type. Record level data are described this way. On the other hand, an object class may be an **individual concept**. This happens when the set of objects corresponding to the object class has one member. Examples are concepts corresponding to single objects, such as "the set of persons in the US" or "the set of service sector establishments in Australia". Aggregate data are described this way. Examples of properties are average income or total earnings.

18.4.3 Conformance

There are no specific conformance criteria for this part of ISO/IEC 11179. It is a framework that ties the other parts of ISO/IEC 11179 together. As such, conformance is not an issue for this part. Each of the other parts of ISO/IEC 11179 has its own conformance clause.

18.4.4 Extensions to the ISO/IEC 11179 standard

Although the ISO/IEC 11179 metadata registry is a complex standard comprising several hundreds of pages, there are users that are attempting to extend these standards to meet various challenges. For example the XMDR project (http://xmdr.org/) states its purpose as being: “...concerned with the development of improved standards and technology for storing

and retrieving the semantics of data elements, terminologies, and concept structures in metadata registries".

18.4.5 Examples of ISO/IEC 11179 metadata registries

The following metadata registries state that they follow ISO/IEC 11179 guidelines although there have been no formal third party tests developed to test for metadata registry compliance.

- Australian Institute of Health and Welfare - Metadata Online Registry (METeOR) - http://meteor.aihw.gov.au/content/index.phtml/itemId/181162
- US Department of Justice - Global Justice XML Data Model (GJXDM) - http://justicexml.gtri.gatech.edu/
- US Environmental Protection Agency - Environmental Data Registry - http://www.epa.gov/edr/
- US Health Information Knowledgebase (USHIK)which ia a health metadata registry - http://www.ushik.org/registry/x/
- US National Cancer Institute - Cancer Data Standards Repository (caDSR) - http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr
- US National Information Exchange Model NIEM which NIEM builds on the demonstrated success of the Global Justice XML Data Model - http://www.niem.gov/

18.4.6 Metadata registry vendor tools that claim ISO/IEC 11179 compliance

A number of commercial tools claim their compliance to ISO/IEC 11179. Some of these are:

- Data Foundations Metadata Registry - http://www.datafoundations.com/solutions/data_registries.shtml
- Oracle Enterprise Metadata Manager (EMM) - http://www.oracle.com/consulting/technology/collateral/integration_build_metamanage.pdf

It should, nevertheless, be noted that there are no independent agencies that certify ISO/IEC 11179 compliance.

18.5 Metadata publishing

Metadata publishing is the process of making metadata data elements available to external users, both people and machines using a formal review process and a commitment to change control processes.

Metadata publishing is the foundation upon which advanced distributed computing functions are being built. But like building foundations, care must be taken in metadata publishing systems to ensure the structural integrity of the systems built on top of them.

18.5.1 Metadata publishing formats

A number of formats exist for the publishing of meta-data. The most important of these are:

- Web Ontology Language (OWL) - used by metadata search engines such as Swoogle (http://en.wikipedia.org/wiki/Swoogle)
- XML Metadata Interchange (XMI), an OMG standard for exchanging metadata
- Common Warehouse Metamodel (CMW), an OMG standard for data warehouse metadata
- KM3 or Kernel MetaMetaModel as used in the Metamodel Zoos. The AtlanticZoo is an open source library of more than 100 metamodels under EPL License. KM3 is a simple Domain Specific Language for specifying metamodels. A number of transformations are available to translate from KM3 to other notations like XMI.

18.6 Metadata discovery and matching algorithms

Metadata discovery is the process of using automated tools to discover the semantics of a data element in data sets. This process usually ends with a set of mappings between the data source elements and a centralized metadata registry.

There are distinct categories of automated metadata discovery:

18.6.1 Lexical Matching

- **Exact match** - where data element linkages are made based on the exact name of a column in a database, the name of an XML element or a label on a screen. For example if a database column has the name "PersonBirthDate" and a data element in a metadata registry also has the name "PersonBirthDate", automated tools can infer that the column of a database has the same semantics (meaning) as the data element in the metadata registry.
- **Synonym match** - where the discovery tool in not just given a single name but a set of synonym.
- **Pattern match** - in this case the tools is given a set of lexical patterns that it can match. For example the tools may search for "*gender*" or "*sex*"

18.6.2 Semantic Matching

Semantic matching attempts to use semantics to associate target data with registered data elements.

- **Semantic Similarity** - In this algorithm that relies on a database of word conceptual nearness is used. For example the WordNet system can rank how close words are conceptually to each other. For example the terms "Person", "Individual" and "Human" may be highly similar concepts.

18.6.3 Statistical Matching

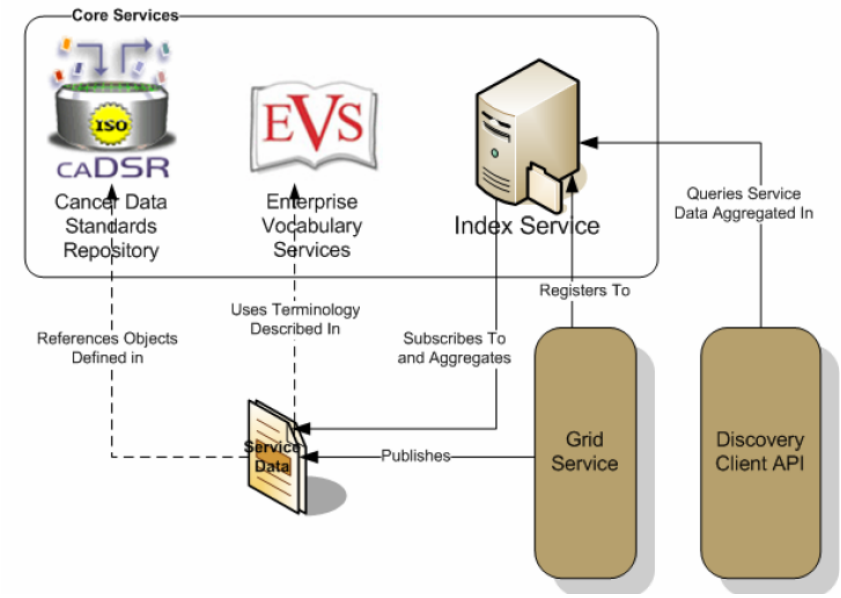
Statistical matching uses statistics about data sources data itself to derive similarities with registered data elements.

- **Distinct Value Analysis** - By analyzing all the distinct values in a column the similarity to a registered data element may be made. For example if a column only has two distinct values of 'male' and 'female' this could be mapped to 'PersonGenderCode'.

- **Data distribution analysis** - By analyzing the distribution of values within a single column and comparing this distribution with known data elements a semantic linkage could be inferred.

18.7 Semantic Service Discovery in the caBIG

A critical requirement of caBIG's infrastructure is that it supports the ability of researchers to discover the available resources. CaGrid enables this ability by taking advantage of the rich structural and semantic descriptions of data models and services that are available. The overall architecture for service advertisement and discovery is shown in the figure below.



Each service is required to describe itself using caGrid standard service metadata. When a Grid service is connected to the caBIG Grid, it registers its availability and service metadata with a central indexing registry service (the GT's Index Service). This service can be thought of as the "yellow pages" and "white pages" of caBIG. A researcher can then discover services of interest by looking them up in this registry using high-level APIs and user applications.

caGrid provides standards for service metadata to which all services must adhere. The basic metadata supported is the Common Service Metadata standard that every service in caBIG is required to provide. This metadata contains information about the service-providing cancer center, such as the point of contact and the institution's name. Extending beyond this generic metadata there are two standards that are specialized depending on whether a data or

analytical service is described. The Data Service Metadata details the domain model from which the Objects being exposed by the service are drawn. Additionally, the definitions of the Objects themselves are described in terms of their underlying concepts, attributes, and associations to other Objects being exposed. Similarly, the Analytical Service Metadata details the Objects using the same format as the Data Service Metadata. In addition to detailing the Objects definitions, the Analytical Service Metadata defines the operations the service provides. The input parameters and output of the operations are defined by referencing the appropriate Object definition. In this way, both the data and analytical services fully define the domain objects they expose by referencing the data model registered in caDSR, and identify their underlying semantic concepts by referencing the information in EVS.

The caGrid discovery API and tools allow researchers to query the Index Service for services satisfying a query over the service metadata. That is, researchers can lookup services in the registry using any of the information used to describe the services. For instance, all services from a given cancer centre can be located, data services exposing a certain domain model or objects based on a given semantic concept can be discovered, as can analytical services that provide operations that take a given concept as input.

18.8 Semantic Service Discovery in the myGrid Project

In MyGrid's Taverna Workbench services can be annotated with semantic descriptions based on ontologies and later discovered based on these descriptions. The myGrid ontology has been described in [my-ont] and is expressed in DAML+OIL (a predecessor of OWL). The reasoning capabilities of DAML+OIL were used in order to have an automated way of developing ontologies since the manual creation and maintenance of classifications was considered a difficult and complex task. DAML-S, which is now known as OWL-S, was subsequently used to describe services and their functionality. The unit of functionality that is described is the *operation* while the service entity is considered as a unit of publication of operations. For each operation the input and output parameters are defined providing a name, a description, a semantic type (for example, a DNA sequence), a format type (e.g. GAME or AGAVE XML format), a transport type (e.g. string), etc. Moreover, each operation is annotated by the overall task being performed (e.g. sequence aligning), the method (i.e. algorithm, for example BLAST) used, the application or tool that implements the functionality and any additional resources it uses, such as a protein database. On the other hand a service description encompasses the provider's description, the author's name, the service's WSDL or URL and so on. Much of the information needed to describe a service and especially its operations is controlled by domain ontologies and vocabularies. In myGrid a suite of ontologies are used that describe molecular biology and bioinformatics concepts as well as organizational, service and workflow concepts.

In order to semantically annotate a given service, the Taverna Workbench integrates with the PeDRo tool (http://pedrodownload.man.ac.uk/). The PeDRo tool provides a graphical interface through which a user is guided to fill the missing semantic information and build XML descriptions of its services using the myGrid ontology suite. These XML descriptions can then be published to a WebDAV server and also advertised to a UDDI registry. On the other hand the Taverna Workbench offers also the ability for the semantic discovery of services through the Feta component [feta]. Feta is composed of two sub-components, the Feta Client GUI that is the user interface for the formulation of semantic queries, and the Feta Engine which is a web service responsible for searching service descriptions that match user's search criteria. The Feta client side GUI is currently able to formulate a number of canned queries such as:

- Find an operation that accepts input of semantic type X or something more general, or find an operation that produces output of semantic type Y or something more specific.
- Find an operation that performs task X or something more specific.
- Find an operation that uses method X or something more specific.
- Find an operation that is function of application/toolkit X or something more specific.

These queries are submitted to the Feta Engine which makes use of RDF-Schema and RDQL for performing the semantic matchmaking. Each service description that was created by PeDRo as described above is retrieved from the UDDI registry and converted to RDF during the initialization phase of the Feta Engine. The use of RDF and RDQL supports complex search queries over the service data but these capabilities are hidden by the Feta Client GUI to make the discovery phase more accessible to the average user. There are plans to integrate Taverna with the GRIMOIRES registry (http://twiki.grimoires.org/bin/view/Grimoires/) that represents all information in RDF. GRIMOIRES is UDDI compliant, supports authentication and access control, has efficient RDF query capabilities with its file-backed, in-memory RDF store, and additionally offers the ability to attach metadata descriptions to UDDI entities (business or services), WSDL operations and messages, or other metadata. The meta information attached can have various forms such as a string, a URI, or RDF data.

18.9 References

[ALP2006] Pinar Alper et al, "S-OGSA as a Reference Architecture for OntoGrid and for the Semantic Grid", GGF16 Semantic Grid Workshop, February 13-16, 2006, Athens, Greece [http://www.semanticGrid.org/GGF/ggf16/papers/OntoGrid-GGF16-SemGrid-Wrkshp.pdf]

[BEC2005] Stefano Beco, Barbara Cantalupo, Ludovico Giammarino, Mike Surridge and Nikolaos Matskanis "*OWL-WS: A Workflow Ontology for Dynamic Grid Service Composition*", accepted to the 1st IEEE International Conference on e-Science and Grid Computing, Dec. 5 - 8, 2005, Melbourne, Australia

[DUR1985] William R. Durrell, Data Administration: A Practical Guide to Data Administration, McGraw-Hill, 1985

[GOB2004] Carole A. Goble and David De Roure. The Semantic Grid: Myth Busting and Bridge Building. In Proceedings of the 16th European Conference on Artificial Intelligence (ECAI-2004), Valencia, Spain, 2004. [http://www.semanticGrid.org/docs/ECAISemanticGrid/ECAISemanticGridFinal.pdf]

[JEF2006] Stephen J. Jeffrey, Jane Hunter "A Semantic Search Engine for the Storage Resource Broker", GGF16 Semantic Grid Workshop, February 13-16, 2006, Athens, Greece [http://www.semanticGrid.org/GGF/ggf16/papers/hunter_jeffrey.pdf]

[KIM1998] Ralph Kimball, The Data Warehouse Lifecycle Toolkit, Wiley, 1998, ISBN 0-471-25547-5

[KOU2006] Manolis Koubarakis, et al, "*Semantic Grid Resource Discovery using DHTs in Atlas*", GGF16 Semantic Grid Workshop, February 13-16, 2006, Athens, Greece [http://www.semanticGrid.org/GGF/ggf16/papers/koubarakis.pdf]

[LOR2005] Phillip Lord, Pinar Alper, Chris Wroe, and Carole Goble *Feta: A light-weight architecture for user oriented semantic service discovery*. In Proceedings of The Semantic Web: Research and Applications: Second European Semantic Web Conference (ESWC 2005), Heraklion, Crete [http://homepages.cs.ncl.ac.uk/phillip.lord/download/publications/european_semantic_web2005_feta.pdf]

[MAR2004] David Martin, et al. "*Bringing semantics to web services: The owl-s approach*". In First International Workshop on Semantic Web Services and Web Process Composition (SWSWPC 2004), Lecture Notes in Computer Science. Springer, July 2004. [http://www.cs.cmu.edu/People/softagents/papers/OWL-S-SWSWPC2004-final.pdf]

[PAH2006] Said Mirza Pahlevi, Isao Kojima "S-MDS: A Semantic Information Service for Advanced Resource Discovery and Monitoring in WS-Resource Framework" GGF16 Semantic Grid Workshop, February 13-16, 2006, Athens, Greece [http://www.semanticGrid.org/GGF/ggf16/papers/said-s-mds.pdf]

[WRO2003] Chris Wroe, et al, "*A suite of DAML+OIL Ontologies to Describe Bioinformatics Web Services and Data*" in International Journal of Cooperative Information Systems special issue on Bioinformatics, March 2003. ISSN: 0218-8430 [http://twiki.myGrid.info/twiki/pub/MyGrid/ServiceOntologies/myGrid_service_ontology_02.pdf]

19 Workflow management systems and standards

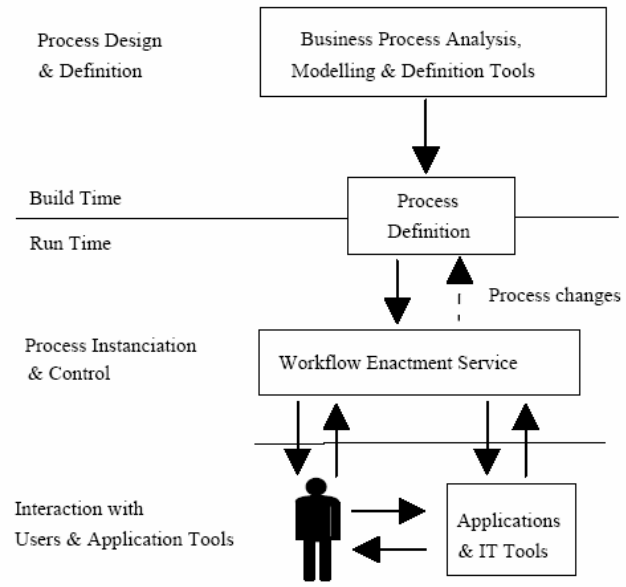
Today computational scientists across all disciplines create ever increasing amounts of often highly complex data. Generated raw and derived data may come from wet lab experiments, large-scale data-intensive and compute-intensive simulations, or real-time observations e.g. from remote sensors.

Technical challenges include not only managing the volume of data, but also the complexity of managing computations distributed over the Grid. In order to support scientists in their data management and analysis tasks, scientific workflows have recently gained increased interest and momentum as a unifying mechanism for handling scientific data. Scientific workflows pose a unique set of challenges due to the special nature of scientific data and the specific needs for large-scale data collection, querying and analysis.

19.1 What is a workflow?

The Workflow Management Coalition [WFMC] defines *workflow* as "The automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules". In other words a workflow consists of all the steps that should be executed in order to deliver an output or achieve a goal. These steps (tasks) can have a variety of complexity and usually are connected in a non-linear way, formulating a directed acyclic graph (DAG). A Workflow Management System defines, manages and executes workflows through the execution of software that is driven by a computer representation of the workflow logic. The description of a workflow includes the definition of different tasks, their interconnection structure and their dependencies and relative order. This description of a workflows operational aspects can be expressed in textual (e.g. XML) or graphical form (e.g. as a graph in Business Process Modelling Notation [BPMN] or Petri nets).

The Workflow Reference Model [WFRM] proposes the following model:



This model defines two major phases for workflows:

- The build phase where the workflow is defined in terms of a textual or graphical language. Various modelling languages that have been proposed include the Web Services Flow Language (**WSFL**), Microsoft's **XLANG** for BizTalk, and Business Process Execution Language for Web Services (**BPEL4WS**) which is the cooperative merging of WSFL and XLANG for Web services orchestration.

- The run phase where the workflow is enacted according to its definition by a workflow execution (enactment) component or a *workflow engine*. At this phase the execution of some tasks may require the interaction with users or other software applications and tools.

More specifically we can identify at least four important aspects of a workflow building and enacting process:

1. User environments, usually graphical, where the user can define a workflow

2. Representation languages that are used to express workflows

3. Translation or compilation of a workflow so that it could be enacted

4. Execution of a workflow and runtime support.

There are cases where all these actors and stages in workflow design and enactment are supported by a (seemingly) single Integrated Development Environment (IDE) that hides the underlying complexity from the user.

19.2 eScience workflows

In addition to the business oriented use cases, workflows have a lot of potential in scientific areas as well. At the current pace of information production there is an unprecedented demand for extraction and processing of knowledge. This is more than evident in various scientific fields such as molecular biology, high energy physics, and astronomy. Consequently, scientific workflows have been proposed as a mechanism for coordinating processes, tools, and people for scientific problem solving purposes [SIN1996]. They aim to support “coarse-granularity, long-lived, complex, heterogeneous, scientific computations”.

In this case however there are some special requirements that differentiate the scientific workflows from the business workflows: they should support large data sets and data flows with a large number of parameterized jobs and the execution is usually done in dynamic environment where resources are not know a priori. Scientists are usually having a hard time trying to locate the data they want, gather them, and find the necessary tools in order to process and analyze them. There are many software tools available to support their scientific experiments but they usually work differently and require a learning phase that’s an impediment to their rapid deployment. Also the different data formats (even if we consider XML formatted documents only) impose either the time consuming task of manual data transformation, or the custom development of wrappers and converters (probably in some scripting programming language, e.g. Perl), which is definitely something beyond a scientist’s area of interest and expertise. In the case of an experiment or study there are also additional issues that relate to the reproducibility of the scenario, the validation and the recording of the provenance of the data inputs. Therefore the composition of the available tools in terms of a scientific workflow in order to orchestrate them for performing some scientific scenario or experiment presents more challenges than in the case of business workflows.

With the advent of Grid technologies the development of workflows technologies that take advantage of the Grid infrastructure and resources has been emerged. In a Grid Computing Environment workflows are concerned with the linkage of constituent services together in order to build either specific Grid applications or even larger, higher level, composite services that can be also accessible through the Grid middleware. Similar to the definition of business workflows given by the Workflow Management Coalition, we could define a Grid workflow as *an automation of a Grid process, in whole or part, during which documents, information or data are passed from one Grid service to another for action, according to a set of procedural rules*.

19.3 Related Projects and Initiatives

There are many European and international efforts that investigate the deployment of workflows in scientific areas in a way that makes capital out of the Grid infrastructure. Some of these efforts are described in the following paragraphs along with their main topic and outcomes.

19.3.1 GridLab

The GridLab project (http://www.Gridlab.org/) aimed at the development of application tools and middleware for Grid environments. GridLab produced a set of application-oriented Grid services and toolkits providing capabilities such as dynamic resource brokering, monitoring, data management, security, information, adaptive services and more. The **Grid Application Toolkit (GAT)** provides applications with access to various GridLab services, resources, specific libraries, tools, etc. in a way that the end-users and especially application developers can build and run applications on the Grid without needing to know details about the runtime environment in advance. Applications use the GAT through a fixed GAT API. Other products of GridLab include **GridSphere**, an open-source portlet based Web portal, **GRMS**, a job scheduling and resource management framework, which allows users to build and deploy job and resource management systems for Grids, and **GAS**, a trusted single logical point for defining security policy for complex Grid infrastructures and Virtual Organizations.

As part of the development of the Work-Flow Application Toolkit (TGAT) the Triana package was extended to include heterogeneous modules executing on remote machines. The enhanced capabilities include the automatic transfer of execution to remote machines, remote program steering using metadata, and two-way interfacing of data-flow and control.

19.3.2 K-Wf GRID

The K-Wf Grid project (http://www.kwfGrid.net/) aims to enable the knowledge-based support of workflow construction and execution in a Grid computing environment. In order to achieve this objective the consortium members will develop a system that will enable users to:

- semi-automatically compose a workflow of Grid services,
- execute the composed workflow application in a Grid computing environment,
- monitor the performance of the Grid infrastructure and the Grid applications,
- analyze the resulting monitoring information,
- capture the knowledge that is contained in the information by means of intelligent agents,
- and finally to reuse the joined knowledge gathered from all participating users in a collaborative way in order to efficiently construct workflows for new Grid applications

The K-Wf project develops the Grid Workflow Description Language (GworkflowDL - http://www.Gridworkflow.org/kwfGrid/gworkflowdl/docs/) that is an XML-based language that makes use of the formalism of Petri nets instead of DAGs in order to describe the dynamic behavior of distributed Grid tasks. The Grid Workflow Execution Service (GWES - http://www.Gridworkflow.org/kwfGrid/gwes/docs/) is the Grid workflow enactment engine of the K-Wf Grid system. GWES builds on top of Globus Toolkit 4 and it conforms to the WSRF standard.

19.3.3 NextGrid

NextGrid (http://www.nextGrid.org/) is a project funded by the European Commission's IST programme of the 6th Framework Programme (IST-511563). The goal and primary output of NextGrid is to define the architecture that will lead to the emergence of the Next Generation Grid. This will prepare the way for the mainstream use of Grid technologies and their widespread adoption by organisations and individuals from across the business and public domains. In addition to new architectural designs, NextGrid will contribute to the key middleware components, application support mechanisms, know-how and standards that underpin the Next Generation Grid. Of particular importance is the willing to address higher-level issues such as co-location, workflow comparison and composition, adaptive behaviour or cross organisational issues, and non-functional aspects of distributed workflows (performance, privacy, security, availability etc.).

19.3.4 DiscoveryNet

DiscoveryNet (http://www.discovery-on-the.net/) is a collection of software built on top of the UNICORE Grid system for arranging database access and knowledge discovery procedures. DiscoveryNet was begun in 2001 at Imperial College of Science and was funded by the Engineering and Physical Sciences Research Council. DiscoveryNet provides a means of describing workflow between analysis service providers, data owners, and scientists who arrange and execute these workflows [ROW2003]. DiscoveryNet makes use of the OGSI components and protocols as well as its own protocol for workflows, Discovery Process Markup Language (DPML). This language is used for constructing, running, and managing Grid services, as well as recording their history. DPML allows conditional processing, loops, sub-routines, and exception handling and uses XML documents as the basic data quantum and URIs as the method of addressing them. The XML processing instructions and resolution of URI schemes are abstracted from the language so as to be fully extensible.

19.3.5 OpenMolGrid

The main objective of the OpenMolGrid project (http://www.openmolGrid.org) was to provide a unified and extensible information-rich environment for solving molecular design/engineering tasks relevant to chemistry, pharmacy and life sciences [OMG2005]. OpenMolGrid was conceived to exploit the power of Grid Computing to shorten the time to solution for drug discovery, specifically the identification of promising new compounds as potential drug candidates. This was to be achieved by extending the currently used local approach where everything is processed on a local resource to the global dimension by building the OpenMolGrid environment on top of the Grid infrastructure provided by UNICORE (Uniform Interface to Computing Resources - http://unicore.sourceforge.net/).

The workflows supported by OpenMolGrid consisted of a sequence of database operations, data conversions and computational steps for the molecular engineering process.

19.3.6 MyGrid

MyGrid (http://unicore.sourceforge.net/) is a UK e-Science pilot project funded by the Engineering and Physical Sciences Research Council. Its objective is the development of the necessary infrastructural middleware (e.g. provenance, service discovery, workflow enactment, change notification & personalisation) that operates over an existing Web services & Grid infrastructure to support scientists in making use of complex distributed resources.

The most important outcome of the myGrid project is the Taverna Workbench and its associated tools. Taverna is a GUI used for assembling, adapting and running workflows enacted by the Freefluo workflow engine. The workflows enacted could incorporate a wide assortment of services and tools. While any web service can be integrated, legacy applications are also supported using the Soaplab-Gowlab wrapper tools.

The Taverna suite is described in more detail in a following paragraph.

19.4 Available Open Source Tools

19.4.1 Taverna

Version Reviewed: 1.3.1

19.4.1.1 Description

The Taverna Workbench (http://taverna.sourceforge.net) allows users to construct complex analysis workflows in the bioinformatics field from components located on both remote (Web Services) and local machines (Java resources), run these workflows on their own data and visualize the results. As a component of the myGrid project, Taverna is available freely under the terms of the GNU Lesser General Public License (LGPL). Effectively Taverna allows a biologist or bioinformatician with limited computing background to construct highly complex analyses over public and private data and computational resources, all from his own PC [TAVOIN].

19.4.1.2 Features

19.4.1.2.1 Workflow Language

The workflow language used in Taverna is called SCUFL. SCUFL stands for 'Simple Conceptual Unified Flow Language'. This language is essentially a data flow centric language, defining a graph of data interactions between different services. It is based on IBM's WSFL, including support for specifying control and data flow. The reasons for developing a proprietary workflow language for Taverna according to [TAVOIN] are historical and conceptual: when the project started there wasn't a well supported and open source language while WSFL didn't seem capable for supporting dataflow in silico experiments.

19.4.1.2.2 User Interface

Taverna provides a data-model centric GUI that allows locating Web Services, composing workflows, running these workflows and viewing and storing their results [TAVMAN]. All steps are graphically supported; the user does not need to have any informatics skills or to deal with a programming language to compose workflows.

The main views of Taverna are:

19.4.1.2.2.1 Advanced model explorer

The Advanced Model Explorer (AME) is the primary editing component within Taverna. Through it any property of a workflow can be loaded, saved and edited.

It presents the various services or processors present in the workflow, the inputs and outputs and the data connections. Within this view the user is able to compose a workflow from different services. For example he is able to create workflow inputs and outputs, connect them to the services, to define temporal constraints (e.g. a processor should only run, when another processor has completed successfully) or to determine alternate processors (if one processor fails another can be launched).

19.4.1.2.2.2 Workflow diagram

This component provides a read only configurable view of the workflow in graphical form. Workflow inputs, outputs and processors appear as coloured boxes with arrows between them to represent data and control links. Options are available to configure the layout and level of detail shown and the resultant diagram may be saved to disc in a number of different formats. It is intended to allow a quick visual overview of the structure of a workflow. In the workflow diagram it is not possible to directly interact with the graphical view, however the current version of Taverna integrates a beta version of a view called “workflow editor” that provides most of the functionality that can be done with the AME by directly interacting with the graphical workflow.

19.4.1.2.2.3 Available Services

This view has the purpose to manage the various services available to the workflow designer and allows the user to manage service libraries, create instances of a service in the form of a processor (can be accomplished easily through drag and drop) within the workflow, search services and other such functionality.

19.4.1.2.2.4 Enactor Invocation

The enactor panel enables the user to enact workflows and view the generated results. It is also possible to store the results.

19.4.1.2.3 Web Services Support

Processors, the main components of workflows, can be as well composed from web services as from local Java resources.

19.4.1.2.3.1 Service discovery

Taverna uses a variety of different mechanisms for discovery of services, ranging from lightweight schemes to heavyweight registries.

Taverna supports:

- Public registries such as UDDI
- GRIMOIRES, an enriched prototype UDDI registry service developed by myGrid, with the ability to store semantic metadata about services

- URL submission, users can add new services, by directly pointing to a URL containing WSDL files

- Processor-specific mechanism, many of the service types provide their own methods for service discovery. E.g. Soaplab (http://www.ebi.ac.uk/soaplab/) and Seqhound (http://www.blueprint.org/seqhound/) services provide mechanisms to allow introspection over an installation or BioMOBY provides its own central registry.

- Scavenging, local disks are scavenged for WSDL files that are introspected over, or users create a web page containing links to service descriptions and, when pointed at the page, Taverna explores all available service descriptions

19.4.1.2.3.2 Service selection

Once the services are gathered by the described techniques they are shown in the “available services” view. Here the services are grouped according to the service locations, which means that services of the same location are grouped together and colour -coded. The user can browse through this services and search them by name. A search based on the properties and semantics of the services is not possible in this view.

That is a problem since Taverna now provides routine access to over 1000 services. Semantic querying would be essential to allow biologist to deal with this number of services. Therefore FETA was integrated into Taverna.

FETA – semantic service discovery

FETA is a plugin for Taverna that allows semantic service search. With the help of Feta it is possible to find services that have inputs or outputs of specified types, perform specified tasks, that are of specified types (e.g. Soaplab services) or whose name or description contains a specified phrase. To allow searching for services using FETA, the services have to be annotated using a tool called Pedro. Pedro allows generating FETA compliant XML based service descriptions that can incorporate different ontologies. Pedro and Feta are components of the MyGrid project that also provides an ontology to annotate web services.

19.4.1.3 Conclusion/Evaluation

With Taverna it is very easy to compose services once they are located into a workflow. Even for non expert users it is very intuitive to compose workflows due to a good graphical support. Taverna allows generating different graphical views of a workflow, showing more or less details. So the user does not lose the overview even when working with very complex workflows.

Taverna does not enforce a common type system. It provides a variety of different techniques to integrate services with all kind of different data types. This is on the one hand an advantage since everybody can quickly add all kind of services together into workflows without the need to know of any specialist programming techniques. On the other hand the big disadvantage is that input and output data formats of the services used by Taverna are mostly semi-structured and heterogeneous. There are a large number of different data formats. These are rarely encoded in XML and there is usually no formal specification.

Often the output data type of one service does not fit with the input data type of the service that has to further process the data. Therefore it is difficult to use services provided by different sites in the same workflow. Taverna provides solutions for this problem, e.g. so

called “shim-services” which are able to transform data. But for creating such services programming skills are required and so they are difficult to use for biologists.

There are currently no mechanisms that checks if the workflow is sensible (e.g. the output data type of one service fits to the input data type of the next service). Workflows that are not sensible can lead to workflow failures. If a workflow fails due to the failure of one service the user can see which service failed and gets the error message generated from this service. These error messages are mostly not understandable to biological users.

The workflow environment does not “understand” the data and so can not perform necessary semantic integration of the data outcomes of those services. So the biologist often needs a significant time to analyse the large amount of often fragmented results.

Service oriented architecture of Taverna has the advantage that the user does not have to install tools and databases locally. On the other hand the disadvantage is that the services can be unreliable and poorly described and therefore difficult to find and to use.

It seems that Taverna is very well suited for the use in ACGT mainly due to the facts that it is open source, provides a very easy usability, is able to incorporate all kind of services, is not bound to a specific data format, and provides a very good extendibility.

The ACGT team can help in finding the best solution for semantic service discovery, semantic integration of the services, and solving the compatibility problems due to the different data formats used by the different service providers. Taverna provides a very good starting point to explore and solve these problems.

19.4.2 Triana

Version reviewed: 3.2.1

19.4.2.1 Description

Triana (http://www.trianacode.org/), a Java-based application distributed under the Cardiff Triana Project Software License (based on the Apache Software License Version 1.1), has been developed at Cardiff University as part of the GridLab (http://www.Gridlab.org/) and GridOneD (http://www.Gridoned.org/) projects.

From a high-level point of view Triana consists of two distinct components-the graphical workflow editor for composing workflows and the workflow manager (also engine) for executing workflows [TRMAN]. Workflows are executed by the workflow manager residing either on the user’s client machine or on a dedicated manager machine in the Grid. For this purpose the Triana workflow editor produces a Java object representing the visual workflow, which is executed by the manager. The manager can be launched on any machine where an appropriate Java Virtual Machine (JVM) is installed. Although the execution is performed on a single machine, the tasks can be distributed, while using the execution machine as central synchronization manager. Triana’s workflow manager is completely independent of the Triana workflow editor; it is self-contained and needs no additional software in order to execute pre-defined flows. However, it is not possible to sign-on and off from the manager during remote execution of such workflows. This means that while remote execution is still possible, the current version of Triana requires the client machine to be permanently connected to the remote machine executing the respective workflow. According to the developers of Triana remote logoff and login during execution was a feature in previous versions, which has been broken in favour of some other functionality. It is currently unclear whether this feature will be fixed in the near future.

Triana is being used by FhG-AIS in the IST-project DataMiningGrid (http://www.dataminingGrid.org/).

19.4.2.2 Features

Specific features of relevance to ACGT include:

- ***Workflow Language -*** No standardized workflow language is used.
- ***User Interface -*** The figure below depicts the workflow editor. On the left side all available units[14] are displayed in a tree-like view. While a lot of units belong to packages already included in the downloadable distribution of Triana, it is also possible to develop units for specific purposes or projects, of course. The central panel views the workspace, in which the workflow is created by dragging the respective units from the tree-view and dropping them onto the workspace. Triana allows opening multiple workflows simultaneously, albeit only the one currently having the focus is executed. A unit's properties are accessed by double-clicking the unit in the workflow (here sqlQueryStatement). Triana provides the following important features for intuitive operation of the editor:
  - **Ability for graphical creation, loading, and storing of workflows:** Workflows are created graphically and can be loaded from and stored to local disk including all parameters entered by the user.
  - **Drag and drop:** Units can be dragged from the tree view on the left hand side and dropped onto the active workspace in an intuitive manner.
  - **Typed input and output of units:** While implementing units, developers can define custom data types passed from one unit to another. During creation of the workflow the editor checks for mismatches and denies connecting units if their input/output types do not match. This feature prevents users from accidentally connecting units in a wrong way, thus creating broken workflows, which will not run. Additionally to creating custom data types, developers may also choose from a wide variety of types already included in Triana.
  - **Search for packages and units by name:** Units are grouped into packages (e.g., DMG/OGSA-DAI) in the tree view. If the unit's name is known, user can search for units, thus alleviating the process of finding a particular unit. This feature also works for partial names and is case-insensitive.
  - **Creation of groups consisting of several units:** Triana provides several ways of combining multiple small units, which form a sub-graph in the workflow, into a single group. These groups behave like "normal units" (e.g., access to the units' parameters by double-clicking the group-unit) and are clearly distinguishable. Each group can be saved to disk and provided to end users by the partners, without editing any code. End users can profit from this feature by using a single coarser grained group, hiding the internal sub-workflow, instead of several fine-grained units. In this way it is possible to provide single but powerful group-units, while at the same time avoiding the

[14] In Triana each node appearing in the workflow, indicating an operation to be carried out, is called a unit. Units are Java classes, which typically issue calls to client-side APIs. Thus, units represent thin wrappers around these client-side classes.

development of units which are too specialized to be used in any other scenario. This is applicable for several sub-workflows, such as searching for and selecting data mining applications from a Grid-wide repository and performing standard data base queries with OGSA-DAI. The latter is depicted in the figure below, which shows exactly the same workflow as the previous figure only with the OGSA-DAI units combined to a single group (yellow).

Additionally, by providing such groups the visible number of nodes required to execute a single job is reduced, which helps users especially during their first steps with the system. Furthermore, grouping units and storing these groups can be done graphically from within Triana, thus enabling also the end-user to create his own customized groups for common tasks.

- **Zooming in and out of workflows:** Triana offers a mechanism to zoom in and out of workflows. This is especially important when creating workflows with a large number of units.

- **Development of new units:** Triana offers a graphical wizard for creating new units to perform custom operations in a workflow. This wizard includes specification of input/output types and creation of a simple graphical panel for displaying the unit's properties, if appropriate and required. More sophisticated panels can be included by providing the full class name of the panel. After all specifications are complete, the wizard creates a code skeleton containing all required standard methods. This skeleton can then be filled with logic using any Java IDE. It is also possible to change and compile units from inside Triana using a built-in editor, which is very useful if only small changes have to be made to a unit.

- ***Web Services Support -*** Triana is capable of discovering Web services either through UDDI or by accessing a user specified URL. Once a Web service is discovered Triana binds to it by analyzing its WSDL description and creating appropriate client-side classes. All messages are sent as via SOAP over HTTP. Whether Triana is also capable of using HTTPS in this context is unclear. Complex data types are supported by generic units, which analyze the WSDL and offer graphical panels according to the content of the respective data types. Therefore, it is not necessary to develop specialized units for any complex data type passed to or from a service.

- ***Grid services support -*** **Discovery of and binding to WSRF compatible services:** Triana is capable of discovering and binding to WSRF-compatible services. This enables users to include applications that are wrapped as services rather than batch-oriented applications started remotely from the command line. When binding to a service, each of its public methods is displayed as a single unit, as depicted in the figure below. These units can then be used to graphically compile a workflow. Triana also provides two units, WSTypeGen and WSTypeViewer, which process the respective service's WSDL file and dynamically renders respective input/output fields (in this example just an int as input/output).

  However, WSRF binding is not perfect just yet. Currently, according to the developers, the implementation of Triana's WSRF binding assumes that creation and manipulation of resources are implemented in the same service rather than using a factory services for resource creation and a second service manipulation of this resource. As the latter is the standard way propagated in most WSRF tutorials, it has to be anticipated that most service developers will implement their services

according to this pattern. The developers of Triana are aware of this problem and seem willing to solve this problem fast.

- **Inclusion of the Grid Application Toolkit (GAT):** The GridLab GAT is a generic and flexible API for accessing Grid services, such as job submission and file movement. It has an adaptor based pluggable architecture that allows different service bindings to be utilized. For example, a GRAM adaptor allows job submission using Globus GRAM, while a GRMS adaptor allows job submission using GridLab GRMS. New adaptors can be developed and plugged in without changing application using the GAT (i.e. Triana).

  Triana uses the GridLab GAT to allow job submission with a workflow, and also to transfer input files to and output files from a remote job.

19.4.2.3 Conclusion/Evaluation

FhG-AIS has great and mostly positive experience with Triana, which has been chosen as workflow editor for the DataMiningGrid project.

Although Triana is not specialized on any problem domain, as for instance Taverna for biological and bio-medical problems, it is a viable option for any project requiring a graphical editor for composing workflows. Triana is very easy to use and all functionality can be accessed graphically via menus in an intuitive way (in contrast to Taverna). For performing custom operations in workflows new units can be created very easily thanks to a wizard, which creates the code skeleton, and to the comprehensive documentation. Its Web services and WSRF capabilities as well as its coupling with GAT make Triana a viable option for almost any Grid project, although the WSRF binding still needs some development. Last but not least, Triana is a very lively project, with dedicated developers, both reflected especially by the mailing lists.

19.4.3 Pegasys

Version reviewed: 0.6

19.4.3.1 Description

Pegasys (http://www.bioinformatics.ubc.ca/pegasys/) is a modular and customizable framework for biological sequence analysis developed by the UBC Bioinformatics Centre under the GNU General Public License. The sequence analysis tasks that are supported include pair-wise and multiple sequence alignment, gene prediction, prediction of RNA sequences and eukaryotic splice, and masking of repetitive elements [SHA2004]. In Pegasys a workflow consists of a set of analyses a biologist wishes to perform on a single sequence or set of sequences. Each task of the workflow is a biological sequence analysis and each such task can accept input from one or more other tasks. Also analyses that are not serially dependent can be executed in parallel.

19.4.3.2 Features

The architecture of the Pegasys is based on a client/server model. The client has a graphical user interface for the creation of workflows and all workflows are sent to the server for execution. The server architecture follows a layered architecture where different layers are responsible for job scheduling, execution, database interaction, and adaptors.

The application layer converts the work flow rendered in XML into a directed acyclic graph (DAG) of analyses in memory. While traversing the DAG, the application schedules all of the analyses on a distributed compute cluster and facilitates the flow of data so that a particular node's program is only executed once all of its inputs are ready (i.e. all of the 'parent' analyses are complete). As each analysis completes, the results are inserted into the backend database layer. Complete reports and computational features of a sequence are inserted into relational tables. The data is exported from the system via the adaptor layer in various formats (currently GFF, GAME XML and raw output from each analysis tool are supported) for human interpretation or for import into other applications.

The main features supported are:

- ***Workflow Language -*** Pegasys uses a proprietary workflow language in XML that is not documented in the user or the developer manuals.
- ***Web Services Support -*** Pegasys does not have any support for web services yet. There are plans on adding the capability of integrating remote services but this has not been implemented yet.
- ***Grid services support -*** The execution of tasks is taking care either by the OpenPBS (http://www.openpbs.org/) “portable batch system”, which seems to be obsolete at the time being, or the Sun Grid Engine (http://Gridengine.sunsource.net/). There is currently no integration with OGSA environments and tools, such as Globus Toolkit, and other service oriented Grid computing middleware.

19.4.3.3 Conclusion/Evaluation

Pegasys presents an intuitive working environment for a biologist to implement sequence analysis scenarios. Nevertheless the centralized approach that is followed, with a central server processing and executing the workflows, in addition to the lack of support for web services and state of the art Grid technologies presents a limited applicability in the ACGT environment.

19.4.4 Kepler

Version reviewed: 1.0 alpha 9

19.4.4.1 Description

Kepler (http://kepler-project.org/) is a workflow system with advanced features for composing scientific workflows useful in a variety of disciplines, e.g., biology, geology, etc. The Kepler project is a crossproject collaboration that includes contributing members from the following projects:

- SEEK: Science Environment for Ecological Knowledge
- SDM Center/SPA: SDM Center/Scientific Process Automation
- Ptolemy-II: Heterogeneous Modelling and Design
- GEON: Cyber-infrastructure for the Geosciences
- EOL: Encyclopaedia of Life

- CIPRes: CyberInfrastructure for Phylogenetic Research
- ROADNet: Real-time Observatories, Applications, and Data Management Network

It is based on the Ptolemy II system (http://ptolemy.eecs.berkeley.edu/ptolemyII/) which is a component assembly framework for the modelling and simulation of concurrent and heterogeneous systems. Ptolemy II supports *multiple models of computation* that describe the interactions and the behaviour of the components. Kepler builds on top of Ptolemy reusing major parts of it (such as its general design and graphical interface) and extending it through various components targeting scientific applications and workflows.

19.4.4.2 Features

Kepler offers some unique advanced features based on the so called actor oriented modeling paradigm of Ptolemy [BER2004]. Each workflow is actually a composition of independent components (actors) that communicate with each other through interfaces called ports. There are two kinds of ports: input ports and output ports, and the input and output ports of different actors are connected via channels. Each actor can be parameterized so that it exhibits a concrete behavior, e.g. a Scale actor that is responsible for scaling its input by a factor can have a parameter to specify this scaling factor. The execution of a workflow is controlled by a director which is the object that orchestrates the actors, initializing them and controlling their execution, and also defining the semantics of the execution that is frequently referred as the "model of computation". For example, in the case of the Process Network (PN) model of computation each actor is an independent thread of control that executes concurrently with the other actors and blocks when trying to read from an empty channel until a message becomes available on it. On the other hand, in a Synchronous Data Flow (SDF) model actors are activated ("fired") when fixed and pre-specified numbers of tokens are available on each of their inputs and each actor produces a fixed number of tokens on each of its outputs. Since each actor declares its consumption and production characteristics the SDF scheduler determines the number of times each actor should be fired and the actor's firing order prior to the workflow execution.

These different models of computation can also be combined in composite workflows where an internal workflow as seen as a composite actor that looks no different from the outside than an atomic actor. Therefore, building hierarchies of workflows, each of them exhibiting its own semantics and model, seems to be very straight forward.

On the sound foundation of Ptolemy II Kepler offers plug-ins and wrappers for working with domain specific (e.g. geology) legacy applications and databases and integration with web services and Grid services.

Specific features include:

- ***Workflow Language -*** Kepler uses Ptolemy's non standard workflow description language called Modeling Markup Language (MoML).
- ***User Interface -*** The user interface is graphical and comes also from Ptolemy. It is called Vergil and using this interface the user can design and execute his/her workflows. There are also some command line tools for executing workflows that are described in MoML.
- ***Web Services Support -*** Kepler was also extended to incorporate web services as actors. The designer of a workflow can supply a URL to the service's WSDL or can give the URL of a service repository (e.g. UDDI or a web page) and Kepler will

"import" all the service descriptions that can be retrieved from there. The WebService actor of Kepler can then be instantiated to any particular operation specified in the web service description (WSDL) and incorporated into a scientific workflow as if it were a local component. In particular, the WSDL-defined inputs and outputs of the service are made explicit via the instantiated actor's input and output ports.

- ***Grid Services Support -*** There are a number of Grid related actors available in the Kepler platform. For example, GridFTP can be used for file transfer. Kepler also includes actors for certificate-based authentication, Grid job submission, and Grid-based data access. Each of these actors access specific Grid-based services using Open Grid Services Architecture (OGSA) interfaces.

19.4.4.3 Conclusion/Evaluation

The Kepler is an advanced workflow management system with some unique features. The areas where we think it leaves something to be desired are the incorporation of state of the art Grid technologies like GT4 and ontology integration. Both of these aspects are currently under development.

19.4.5 VDS - The GriPhyN Virtual Data System

Version reviewed: 1.4.5

19.4.5.1 Description

The Virtual Data System (VDS - http://vds.uchicago.edu/), formerly known as Chimera, was developed as part of the GriPhyN project (http://www.griphyn.org/) which is funded by NSF (Information Technology Research Grant ITR-0086044) and aims to support large scale data management in physics experiments [FOS2002]. It is a set of tools for expressing, executing, and tracking the results of workflows. It includes Pegasus (http://pegasus.isi.edu/) (Planning for Execution in Grids), which is a workflow planner that accepts abstract workflows specified in a high level Virtual Data Language (VDL) and maps them to concrete workflows that can be executed on the Grid. The execution of the workflow is taken care by Condor's DAGMan (http://www.cs.wisc.edu/condor/dagman/).

19.4.5.2 Features

An abstract workflow is a directed acyclic graph (DAG) composed of tasks and data dependencies between them. Each task comprises a logical transformation and has inputs and outputs in the terms of logical filenames. A logical filename represents some data ('virtual data') whose exact physical location is not known or not relevant at the abstract workflow level. In an abstract workflow the transformations (tasks) are not bound to any executable so in a sense they are also abstract and they just define an interface (input and output) for a concrete application or tool. The abstract workflow description language that is used is VDL which comes in two formats: a plain text format and a XML format.

After the creation of an abstract workflow the following step is the generation of the concrete workflow. This phase should produce a workflow that is executable and so it contains the following tasks:

- For each logical file specified in the abstract workflow find its physical location

- Find the physical locations of each transformation along with the preferred locations that will host the execution of the transformation based on its computational requirements
- If the (physical) locations of the data and the tasks that implement the transformations do not match, include additional tasks in the concrete workflow to move data and tasks around

The whole process of generating concrete workflows is called planning and Pegasus is the Grid aware planner component in GriPhyN [DEE2005]. In order to construct a concrete workflow Pegasus uses a number of Globus services like the Replica Location Service (RLS) for finding the physical location of files, the Monitoring and Discovery System (MDS) that provides details about the computational resources (number of processors, available memory, etc, and GridFTP for data transfer. The available tasks to implement the functionality of transformations are searched for in the Transformation Catalog (TC) which also provides information about their performance characteristics. It is also possible that some of the virtual data contained in an abstract workflow is not physically available. In these cases Pegasus consults the additional VDS catalogs in order to find ways to produce them and enhances the concrete workflow with the appropriate tasks.

The Pegasus system tries to make its best on optimizing the execution of workflows and coming up with an efficient execution plan. One such optimization is an attempt to eliminate parts of a workflow that produce data that already exist somewhere in the Grid and could be used. An extreme case of this is, if the needed workflow output is available, the option for not executing the workflow. This optimization is based on the heuristic that it costs more to produce the data by executing a component than to locate and transfer them.

The final outcome of the Pegasus planning process is a concrete workflow represented as a DAG that is submitted to DAGMan or Condor-G for execution. Any data produced are registered in the RLS and VDS catalogs for subsequent execution of workflows. Specific features include:

- ***Workflow Language -*** The abstract workflows are described in the Virtual Data Language (VDL) which comes in two flavors, textual and XML.

- ***User Interface -*** The user interacts with the VDS and Pegasus systems mostly through the command line tools. In the recent releases the Chiron portal is included that presents a web interface for user management, job submission, workflow visualization (via the GraphViz drawing tool), etc accessible through a web browser.

- ***Web Services Support -*** It is unclear if any of the transformations participating in an abstract workflow could be a web service. The Pegasus infrastructure and the VDS catalogs seem to be tailored to support the pre-web service Grid middleware. Nevertheless, the Chiron portal provides a WSDL web service interface for virtual data discovery, composition, and integration.

- ***Grid Services Support -*** Pegasus has as a prerequisite a number of Grid middleware components coming from the Globus Toolkit and Condor. This Grid foundation is distributed as part of the virtual data toolkit (VDT - http://vdt.cs.wisc.edu/index.html) or can be downloaded and installed separately. According to documentation VDS should work with the Globus Toolkit version 4 but as mentioned above we are not sure if it can take advantage of the new WSRF compliant infrastructure.

19.4.5.3 Conclusion/Evaluation

VDS with its virtual data concepts and abstract workflows support appears to be a very interesting system. The abstract specification of workflows gives the user the ability to describe her experiments and scenarios without being concerned with the low level details. Also, the intelligent transformation of an abstract workflow to a concrete one, at run time, gives flexibility and execution efficiency. On the downside the seeming unfriendliness to web services and WSRF compliant Grid services somewhat limits its use in a modern, service oriented Grid environment.

19.4.6 Commodity Grid Kit

Version reviewed: 4.1.4

19.4.6.1 Description

The Commodity Grid (CoG) Kit (http://wiki.cogkit.org/) allows Grid users, Grid application developers, and Grid administrators to use, program, and administer Grids from higher-level framework. The Java CoG Kit enhances the capabilities of the Globus Toolkit by introducing Grid workflows, control flows, and a task based programming model.

19.4.6.2 Features

The Java CoG Kit is actually a client side toolkit that greatly enhances the usability of the Grid layer by providing a common API across different Grid implementation and more specifically the Globus toolkit [LAS2001]. It currently supports the GT2, GT3 and GT4 toolkits. The underlying workflow engine is called Karajan and it facilitates the submission and management of complex tasks in the Grid. In Karajan the workflows are described in a proprietary XML language permits the submission of Grid tasks, along with their dependencies, and the monitoring of their execution. Karajan is based on GridAnt (http://www-unix.globus.org/cog/projects/Gridant/) and supports some additional capabilities, like:

- Control abstractions to support iteration and conditional execution of tasks
- Parallel execution of tasks
- Support of large numbers of parallel tasks through its custom threading support instead of Java's native threads
- Checkpointing and resuming of tasks

Workflows executed by Karajan could also make use of a number of convenience libraries that offer specific functionalities such as a task library to enable access to Grid services, a forms library to enable the dynamic creation of forms as part of workflow tasks, and a Java library to extend the workflow language with elements based on Java classes. Also support for common collection data types such as lists and maps that are specifically targeted to support parameter studies is provided.

Some of its most important features include:

- ***User Interface*** **-** The Java CoG Kit includes various command line tools and a GUI to enact Karajan workflows. A screenshot of it is shown below:
- ***Web Services Support*** **-** Due to its major Grid orientation there seems that no special Web Services support is provided by Karajan outside of the Grid

environment. Nevertheless, the incorporation of the GT 4 handler in Karajan entails the ability to integrate WSRF compliant Web Services.

- ***Grid Services Support -*** Karajan is based on the CoG Kit abstractions so it is able to take advantage and integrate with the Globus Toolkit versions 2, 3, and 4.

19.4.6.3 Conclusion/Evaluation

CoG's ability to formulate workflows in heterogeneous Grid environments, despite the fact that the underlying resources use different versions of Grid services and standards, is often cited as one of its strong points to achieve interoperability and portability.

19.5 References

[WFMC] Workflow Management Coalition, http://www.wfmc.org/

[WFRM] Workflow Reference Model, http://wfmc.org/standards/model.htm

[BPMN] Business Process Modeling Notation http://www.omg.org/cgi-bin/doc?dtc/2006-02-01

[SIN1996] Munindar P. Singh, Mladen A. Vouk "Scientific Workflows: Scientific Computing Meets Transactional Workflows", Position paper in *Reference Papers of the NSF Workshop on Workflow and Process Automation in Information Systems: State-of-the-art and Future Directions*, May 1996. http://www.csc.ncsu.edu/faculty/mpsingh/papers/databases/workflows/sciworkflows.html

[ROW2003] Anthony Rowe, Yike Guo, Dimitrios Kalaitzopoulos, Michelle Osmond, and Moustafa Ghanem *The Discovery Net System for High Throughput Bioinformatics*. ISMB 2003. [http://www.iscb.org/ismb2003/paperAbstracts/btg1031.pdf]

[OMG2005] OpenMolGRID - Open Computing Grid for Molecular Science and Engineering, Final Report, ISBN 3-00-016007-8, July 2005, http://www.fz-juelich.de/nic-series/volume29/volume29.html

[MAR2004] David Martin, Massimo Paolucci, Sheila McIlraith, Mark Burstein, Drew McDermott, Deborah McGuinness, Bijan Parsia, Terry Payne, Marta Sabou, Monika Solanki, Naveen Srinivasan, and Katia Sycara. "*Bringing semantics to web services: The owl-s approach*". In First International Workshop on Semantic Web Services and Web Process Composition (SWSWPC 2004), Lecture Notes in Computer Science. Springer, July 2004. [http://www.cs.cmu.edu/People/softagents/papers/OWL-S-SWSWPC2004-final.pdf]

[LOR2005] Phillip Lord, Pinar Alper, Chris Wroe, and Carole Goble *Feta: A light-weight architecture for user oriented semantic service discovery*. In Proceedings of The Semantic Web: Research and Applications: Second European Semantic Web Conference (ESWC 2005), Heraklion, Crete [http://homepages.cs.ncl.ac.uk/phillip.lord/download/publications/european_semantic_web2005_feta.pdf]

[WRO2003] Chris Wroe, Robert Stevens, Carole Goble, Angus Roberts, Mark Greenwood "*A suite of DAML+OIL Ontologies to Describe Bioinformatics Web Services and Data*" in International Journal of Cooperative Information Systems special issue on Bioinformatics, March 2003. ISSN: 0218-8430 [http://twiki.myGrid.info/twiki/pub/MyGrid/ServiceOntologies/myGrid_service_ontolog

y_02.pdf]

[BEC2005] Stefano Beco, Barbara Cantalupo, Ludovico Giammarino, Mike Surridge and Nikolaos Matskanis "*OWL-WS: A Workflow Ontology for Dynamic Grid Service Composition*", accepted to the 1st IEEE International Conference on e-Science and Grid Computing, Dec. 5 - 8, 2005, Melbourne, Australia

[GOB2004] Carole A. Goble and David De Roure. *The Semantic Grid: Myth Busting and Bridge Building*. In Proceedings of the 16th European Conference on Artificial Intelligence (ECAI-2004), Valencia, Spain, 2004. [http://www.semanticGrid.org/docs/ECAISemanticGrid/ECAISemanticGridFinal.pdf]

[PAH2006] Said Mirza Pahlevi, Isao Kojima "*S-MDS: A Semantic Information Service for Advanced Resource Discovery and Monitoring in WS-Resource Framework*" GGF16 Semantic Grid Workshop, February 13-16, 2006, Athens, Greece [http://www.semanticGrid.org/GGF/ggf16/papers/said-s-mds.pdf]

[ALP2006] Pinar Alper, Oscar Corcho, Ioannis Kotsiopoulos, Paolo Missier, Sean Bechhofer, Dean Kuo, Carole Goble "*S-OGSA as a Reference Architecture for OntoGrid and for the Semantic Grid*", GGF16 Semantic Grid Workshop, February 13-16, 2006, Athens, Greece [http://www.semanticGrid.org/GGF/ggf16/papers/OntoGrid-GGF16-SemGrid-Wrkshp.pdf]

[JEF2006] Stephen J. Jeffrey, Jane Hunter "*A Semantic Search Engine for the Storage Resource Broker*", GGF16 Semantic Grid Workshop, February 13-16, 2006, Athens, Greece [http://www.semanticGrid.org/GGF/ggf16/papers/hunter_jeffrey.pdf]

[KOU2006] Manolis Koubarakis, Zoi Kaoudi, Iris Miliaraki, Matoula Magiridou, Antonios Papadakis-Pesaresi "*Semantic Grid Resource Discovery using DHTs in Atlas*", GGF16 Semantic Grid Workshop, February 13-16, 2006, Athens, Greece [http://www.semanticGrid.org/GGF/ggf16/papers/koubarakis.pdf]

[TAVOIN] Tom Oinn, Mark Greenwood, Matthew Addis, M. Nedim Alpdemir, Justin Ferris, Kevin Glover, Carole Goble, Antoon Goderis, Duncan Hull, Darren Marvin, Peter Li, Phillip Lord, Matthew R. Pocock, Martin Senger, Robert Stevens, Anil Wipat and Chris Wroe. "*Taverna: Lessons in creating a workflow environment for the life sciences*" accepted for publication in Concurrency and Computation: Practice and Experience Grid Workflow Special Issue

[TAVMAN] Taverna User Manual v1.3.1, http://taverna.sourceforge.net/usermanual/manual.html

[TRMAN] Triana User Guide, https://forge.nesc.ac.uk/docman/view.php/33/104/UserGuide.pdf

[SHA2004] Sohrab P Shah, David YM He, Jessica N Sawkins, Jeffrey C Druce, Gerald Quon, Drew Lett, Grace XY Zheng, Tao Xu, BF Francis Ouellette. *Pegasys: software for executing and integrating analyses of biological sequences*. BMC Bioinformatics 2004, 5:40 [http://www.biomedcentral.com/content/pdf/1471-2105-5-40.pdf]

[BER2004] *Kepler: An Extensible System for Design and Execution of Scientific Workflows*, I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludäscher, S. Mock, system demonstration, *16th Intl. Conf. on Scientific and Statistical Database Management (SSDBM'04)*, 21-23 June 2004, Santorini Island, Greece. [http://www.sdsc.edu/~ludaesch/Paper/ssdbm04-kepler.pdf]

[DEE2005] Ewa Deelman, Gurmeet Singh, Mei-Hui Su, James Blythe, Yolanda Gil, Carl Kesselman, Gaurang Mehta, Karan Vahi "*Pegasus: a Framework for Mapping Complex Scientific Workflows onto Distributed Systems*", Submitted to the Scientific Programming Journal,January 2005

[http://pegasus.isi.edu/pegasus/publications/sciprog_submitted.pdf]

[FOS2002] I. Foster, , J. Voeckler, M. Wilde, and Y. Zhao, "Chimera: A Virtual Data System for Representing, Querying, and Automating Data Derivation," Proceedings of Scientific and Statistical Database Management, 2002.[http://www.griphyn.org/documents/document_server/uploaded_documents/doc--156--Chimera_SSDBM_2002.pdf]

[FRE2001] J. Frey, T. Tannenbaum, M. Livny, I. Foster, S. Tuecke. "*Condor-G: A Computation Management Agent for Multi-Institutional Grids.*" Proceedings of the Tenth International Symposium on High Performance Distributed Computing (HPDC-10), IEEE Press, August 2001. [http://www.cs.wisc.edu/condor/doc/condorg-hpdc10.pdf]

[LAS2001] Gregor von Laszewski, Ian Foster, Jarek Gawor, Peter Lane "*A Java Commodity Grid Kit*", Concurrency and Computation: Practice and Experience, 2001, Vol. 13, No. 8-9, pages 643-662, [http://www.mcs.anl.gov/~gregor/papers/vonLaszewski--cog-cpe-final.pdf]

20 3D Visualization and tools for the visual query of data

20.1 Introduction

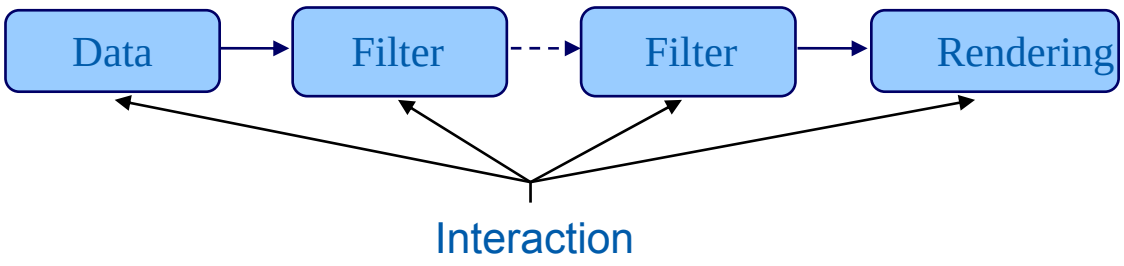
Given the fact that much of the technology that will be used in ACGT is in a state of rapid change, there will be an increasing need for new interactive visualization methods.

Experiences show that interactive 3D visualization can provide a clearer view on complex structures, since they can, in some cases, disclose more structure than 2D graphs. The most spectacular example of 3D visualization is of the interactive visualization of genomics data, in the SARA: SARAgene - an environment to explore genomes in a virtual reality environment (http://www.sara.nl/projects/projects_07_03_eng.html)

In ACGT, the complexity, dimensionality and size of the data to be visualized are expected to increase rapidly. This calls for a flexible and powerful visualization environment. Apart from the needs for the 3D visualization of the results of the Oncosimulator there are needs for such tools to be used for the interactive 3D visualization of scatter plots, Principal Component Analyses, 3D graphs/networks and perhaps even 3D self-organizing maps.

20.2 State of the Art

Many scientific visualization environments are designed around the dataflow paradigm, as illustrated below. In this paradigm, input data is processed by one or more components, configured in a pipeline of data exchanging components that transform the input data into geometrical primitives that can be rendered into an image. In interactive visualization systems, the user is permitted to interact with the individual stages in order to control the parameter settings of each stage.



For interactive visualization systems the responsiveness of the system as a whole depends on several factors:

- throughput of data I/O
- execution time of the visualization filters

- frame rate of the rendering system
- interaction response time

One way to address the performance requirements for an interactive visualization pipeline is by decomposition of the pipeline over distributed resources, like a Grid. Here, the assumption is that components in a visualization pipeline can exploit properties of computational resources to accelerate the execution of each component. Such properties include: proximity to the input data, availability of specialized hardware resources (such as hardware accelerated graphics cards, multi-CPU systems, large-memory systems, fast network links, etc.), software resources (such as specialized algorithms). The distribution of components over distributed systems implies communication between components, which is overhead over the non-distributed pipeline. To be effective, the total gain of distribution must therefore be higher than the communication overhead.

20.3 Interactive visualization … on the Grid?

Although the use of Grid resources for the purpose of scientific visualization has been explored in various projects [SLO2003, BRO2004a], the encapsulation of *interactive* visualization facilities into first-order Grid resources is a relatively new, but active research area [BRO2004b, CHA2004, SHA2003]. In particular, the use of visualization resources in a concerted collection of resources for interactive visualization poses several challenges. First; interaction is a contradictory facility in most of today's Grids. Most computing resources on Grids today are batch systems and therefore do not support interactive modes of use. Second; the overhead associated with the encapsulation of computational resources into Grid services (or Web services as proposed by the "Web Services Resource Framework", or WSRF [CZA2004]) inhibits their application in interactive environments, at least with current implementations.

Because today's Grid infrastructures do not lend themselves for interactive computing, we will develop a coordinating infrastructure based on the concept of "tuple spaces" [CAR1989, GEL1985]. A tuple space is an implementation of the associative memory paradigm for parallel/distributed computing. The characteristics of tuple spaces make them very suitable for our purpose. An open subject that will be investigated in this part of our work is concerned with the integration of resources that have been encapsulated into Grid or Web services into a tuple space architecture.

20.4 Description of the tools to be integrated into ACGT

- ***VTK - The Visualization Toolkit.*** The Visualization ToolKit (VTK[15]) is an open source, freely available software system for 3D computer graphics, image processing, and visualization used by thousands of researchers and developers around the world. VTK consists of a C++ class library, and several interpreted interface layers including Tcl/Tk, Java, and Python. VTK compiles and runs on most UNIX platforms, Windows and MacOS. VTK supports a wide variety of visualization algorithms including scalar, vector, tensor, texture, and volumetric methods.

[15] http://www.vtk.org/

- ***Vtkfly - the UvA visualization environment.*** VTK is not a turn-key visualization package. Instead it provides the building blocks to construct interactive visualization environments. Vtkfly, developed at the University of Amsterdam, uses VTK to encapsulate visualization methods into a user-friendly, flexible, high-performance interactive visualization environment. Vtkfly has been used for the visualization of data from several scientific domains, including medical imaging, astrophysics, flow simulation and biology. Vtkfly supports, among others:
    - interactive 3D visualization of a wide range of input data; multicolumn text files, PDB, DICOM, polygonal data formats, and many more
    - a transparent facility to distribute visualization over distributed systems for optimal interactive response
    - a plug-in architecture to include new visualization methods
    - interactive manipulation of graphical representations
    - interactive methods to perform measurements on 3D visualizations
    - annotation of landmarks in visualizations
    - a scripting interface for the creation of visualizations and animations off-line
    - support for collaborative visualization across distributed sites.

20.5 Visual interfaces to query data model and data.

As specified in document ACGT_D2.1, the ACGT project will handle various kinds of data types and data. Data types are intended to be described by the ACGT master ontology. This ontology and real data pieces (i.e., data that conform to the ACGT ontology) will be stored somehow on distributed data servers accessible through the ACGT Grid infrastructure.

For the point of view of the end users, which are supposed not being specialized in database technologies, ontology and data have to be *accessible* in a *friendly* way. Accessible means here that users will *query* ACGT data bases to retrieve data of interest. Friendly means that the ACGT project should provide easy-to-use interfaces that will hide the complexity of both ontology and real data, as well as hidden the ACGT technical platform.

The following paragraphs will summarize the actual state of the art of systems providing end-users with an access to databases. Since this field of computer science has a long time story, our purpose is not to give an exhaustive list of existing systems. Hence, this paper solely presents the major approaches propose to the end-users to query databases and to visualize the query results, these approaches being illustrated by examples from existing software tools. The idea is to expose the features that could be relevant to design and implement a VQL-based system for ACGT.

20.5.1 End-user access to databases: a state of the art

20.5.1.1 Command-line access

Current database systems, whether they are relational, object oriented or XML based, always provide a text-based query language. Such a language provides the user with a *command-line* access to both the database structure (i.e., the description of the data types stored in the database) and the real data. In that way, some standard languages have been proposed: SQL to query relational databases, OQL to query object oriented databases and

XQuery for XML based databases. Particular versions of these languages, and especially SQL, have also been adapted to target the complex process of simultaneously querying several (possibly distributed) databases (e.g., IBM's DiscoveryLink [HAA2001]).

For the purpose of end users that need to access data repositories, but are not familiar with computer systems, a query language is far from being easy to use. First, there is the difficult step of learning these languages. Moreover, in the field of biomedical information, different types of data (sequences, micro-arrays, images, etc) may be stored in different DBMS, requiring to lean different languages. Second, the data structure (i.e. data types, attributes) has to be known to formulate a query, and that kind of information is not directly accessible at query writing time. Finally, since we are talking about command-line access to databases, query results are frequently presented in poorly-formatted text forms that are rarely straightforward to interpret.

20.5.1.2 Web-based access

To circumvent the problem of using a command-line, databases access has greatly benefited from the advance of the World Wide Web. Now, the end-users can fill in pre-formatted query forms and the results can be nicely formatted to help users interpreting the query results.

Such web-based database accesses are widely used in the biomedical community; see for example the extensively used web portal of the US-based National Centre for Biomedical Information (NCBI, www.ncbi.nlm.nih.gov).

Figure 33: Querying biomedical data on the NCBI web portal (http://www.ncbi.nlm.nih.gov/). This web page snapshot displays the results of searching for all entries published between 1996 and 2006 related to 'glucocerebrosidase'. This page gives a very interesting overview of the results found in the various databases maintained at the NCBI.

If this kind of database access is quite easy to use for the end-users, there is a rapidly emerging lack: a user can only execute particular queries since they are pre-formatted with

web forms that do not give access to the full expressivity power of previously mentioned query languages. Taking the example of the NCBI, a query is usually a boolean expression of text-free keywords possibly decorated with a data type (Figure 33 and Figure 34). However, that system does not give a direct access to the data types at query writing time: one has to read some additional documentation situated elsewhere on their web portal.

Figure 34: Results of the query from Figure 33 that relate to the OMIM database. On such a page a very relevant information, apart a summary of the 5 OMIM entries that relates to our query, is the 'Links' hypertext link located on the right of each summary. Each link allows the user to see other types of data located in other NCBI's databases.

More sophisticated web-based systems exist, such as SRS [ETZ1996] and TAMBIS [STE2000], which provide the possibility to create more complex queries either with HTML forms (Figure 35: SRS query at the EBI web portal (http://srs.ebi.ac.uk)

or using Web browser embedding Java Applet. Usually these forms allow the specification of queries with visually created boolean expressions, the data types being presented to the user (like on Figure 35: SRS query at the EBI web portal (http://srs.ebi.ac.uk)

Query creation is then facilitated, but these more advanced systems still limit the expressivity of queries in comparison with database query languages.

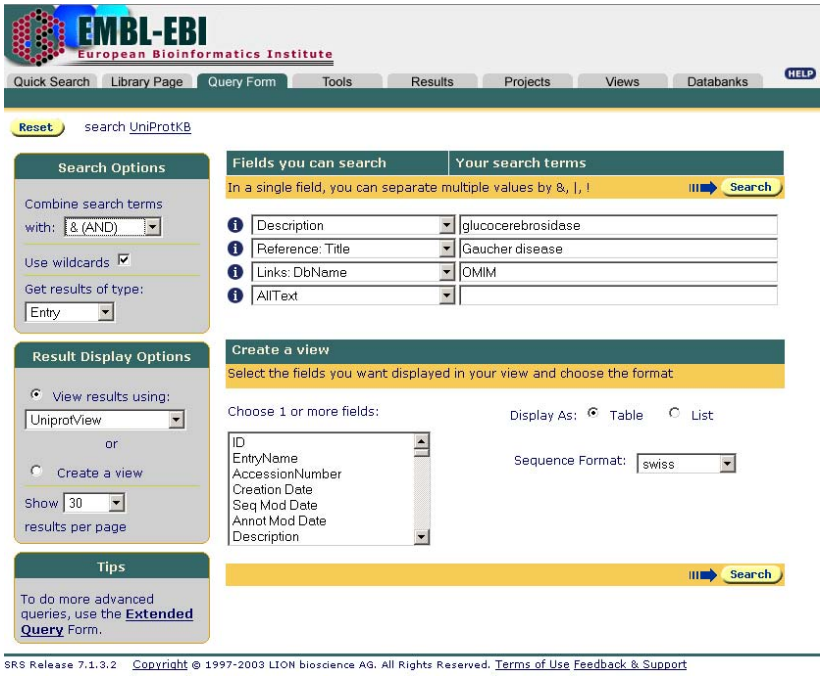


Figure 35: SRS query at the EBI web portal (http://srs.ebi.ac.uk)

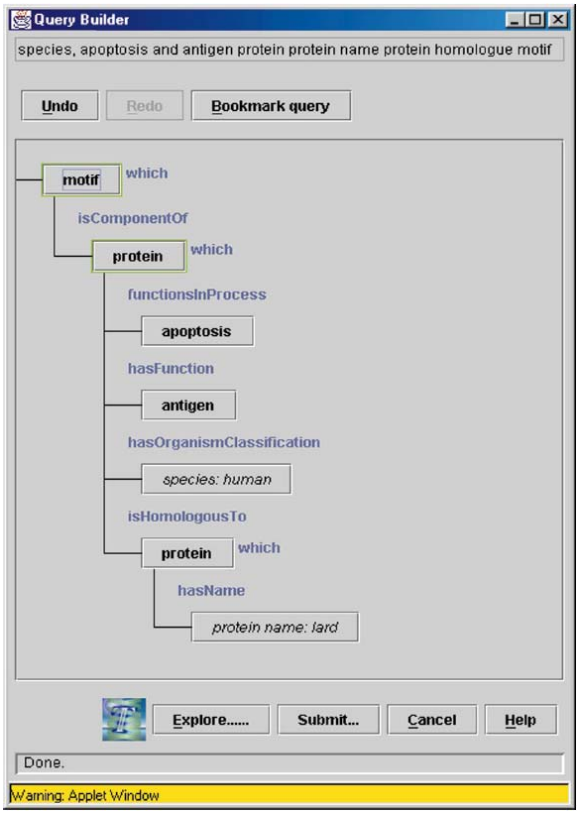


Figure 36: TAMBIS graphical query builder running from MOZILLA web browser

20.5.2 Interactive Graphical User Interface access

Interactive graphical user interface access to database systems relies upon visual query languages (VQLs). Query by Example (QBE, [ZLO1977]) has been the first VQL proposed to query relational databases. Queries are created by 'assembling' visual representation of tables, constraints being added in the columns of the tables. This system has then been extended to provide a more convenient graphical display still available today in database software such as Microsoft Access.

Visual queries described using the graph paradigm is probably the most prominent VQL today available. The graph is used to represent the elements of the schema describing the database structure. Significant systems like GOOD [GEM1993], Hy+ [CON1997], Gql [PAP1994], Hyperlog [POU2001], the system from Butler *et al.* [BUT2005], Snow [WWWNOR], HyperFlow [DOT2005] and GenoLink[DUR2006] provide visual graph query 'languages'.

Such systems are of particular interest for the end users. First, they usually explicitly display the schema model as a unique graph. In that way, the user does not need to know which kind of database he/she targets: whether underlying DBMS is relational or object oriented, whether users target a single database or a set of distributed ones, the visual interface displays the schema in a unique way. The schema graph itself may rely upon a particular data modelling system (see below) providing more flexibility and independence with regard to the DBMS implementation(s) used.

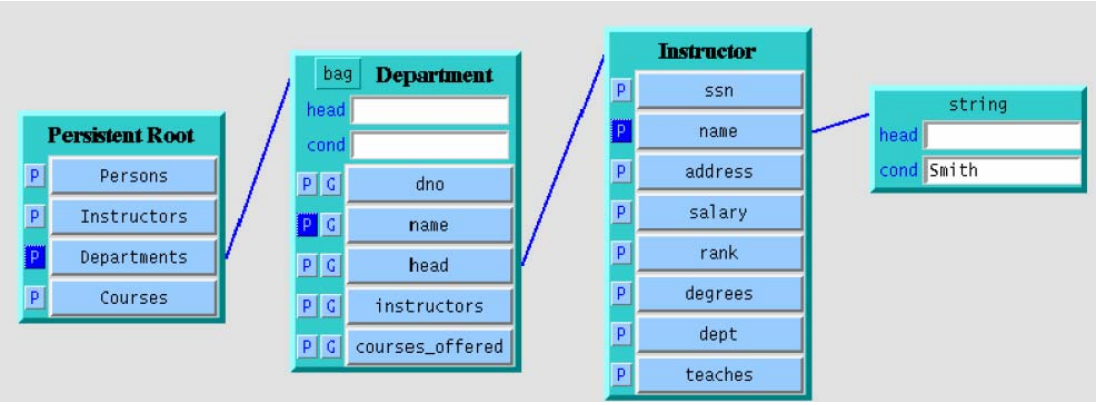


Figure 37: An example of modern Query by Example (from[DOT2006])

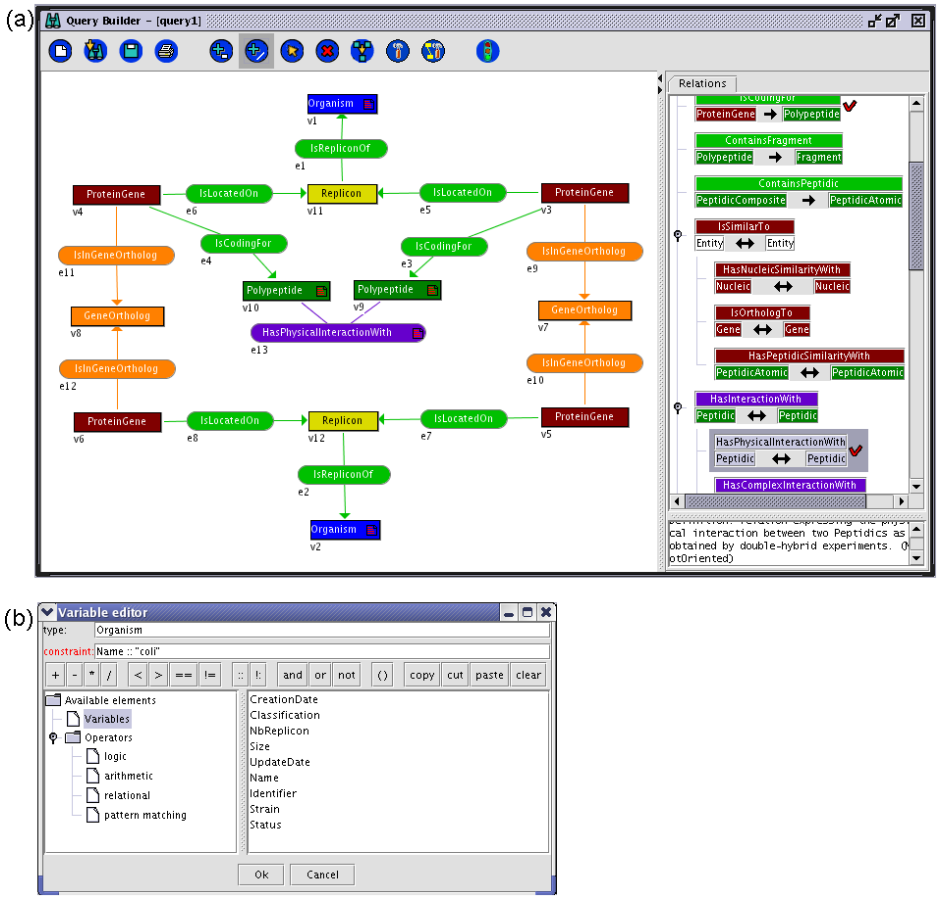


Figure 38: The GenoLink Query Builder. (a) Main window. The left panel displays the graph query being constructed. The right panel displays either the hierarchy of classes or the hierarchy of associations of the data model. Here the user is adding an association therefore the hierarchy of associations is shown. The associations with non empty set of instances are marked with a red "V", allowing the user to quickly know data types having real instances in the database. (b) Clicking on a vertex or edge will popup this constraint editor to add an algebraic constraint on the corresponding object. Here the name of the organism (represented by vertex v2) should match "coli".

Second, both the query and the results are displayed as graph facilitating the interpretation of results since they are written like the query. GenoLink goes one step further by allowing the users to visually explore the neighbours of vertices reported in the results, see following figure.

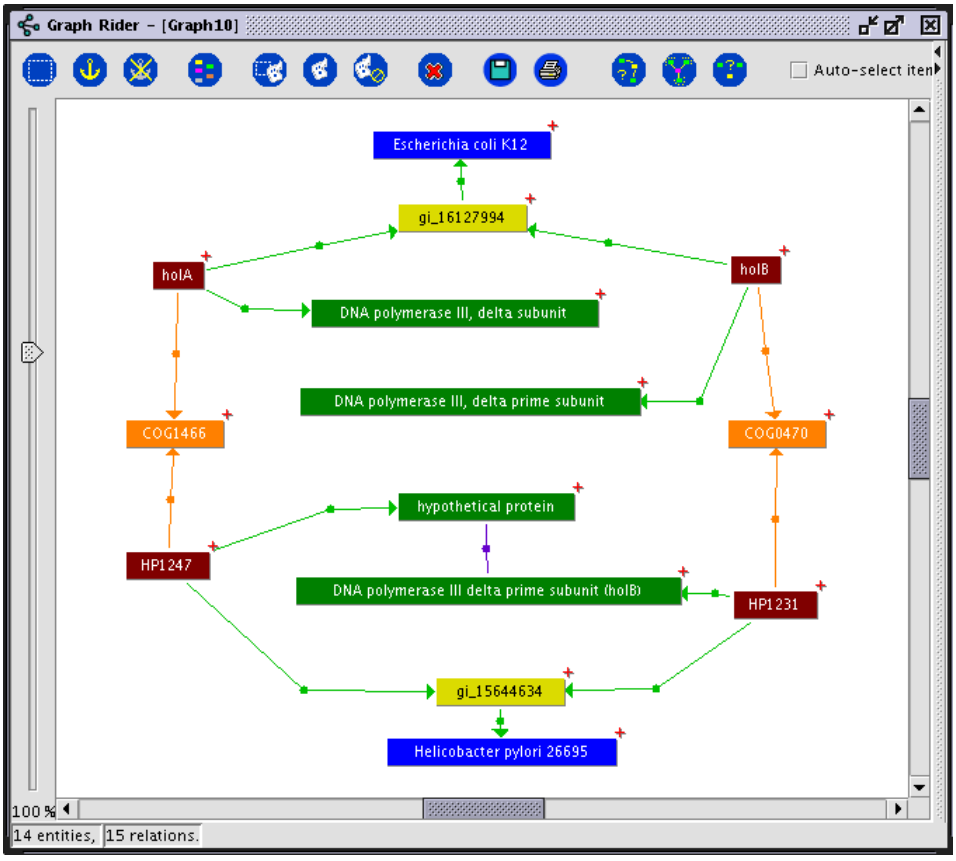


Figure 39: GenoLink Result Graph Explorer. This snapshot shows an example of a result graph corresponding to the Query from Figure 38. The edge linking the two H. pylori Polypeptides corresponds to a physical interaction. The red crosshair on the top-right of some vertices denotes that they are linked to some others that are not currently shown. These vertices may therefore be further expanded to gain more information about the full data graph. In this example, this operation has been performed on vertices holA and holB (from E. coli) in order to display the corresponding Polypeptides (DNA polymerase III) that were not part of the query (see Figure 38).

Third, systems like HyperFlow and GenoLink, relies upon an intermediate data modelling system capable of representing complex data schema. This additional level of abstraction implies that HyperFlow and GenoLink do not rely on a particular DBMS implementation. HyperFlow relies on OWL, whereas GenoLink relies on an entity-relationship knowledge representation system (AROM, [GEN2000]). Vertices and edges of their query graphs are then tidily linked to OWL or AROM entities. Now, to execute a query against a real database (where the data is actually stored), the graph query has to be translated and passed to the DBMS for execution. This step is not yet implemented in HyperFlow. GenoLink uses a different approach since it implements its own graph query engine, the DBMS being only used to feed that graph query engine with real data.

Finally, a system like HyperFlow combines a visual query language with a visual scientific workflows builder, thus providing a single graphical user interface capable of creating complex ‘queries’ in an easy way.

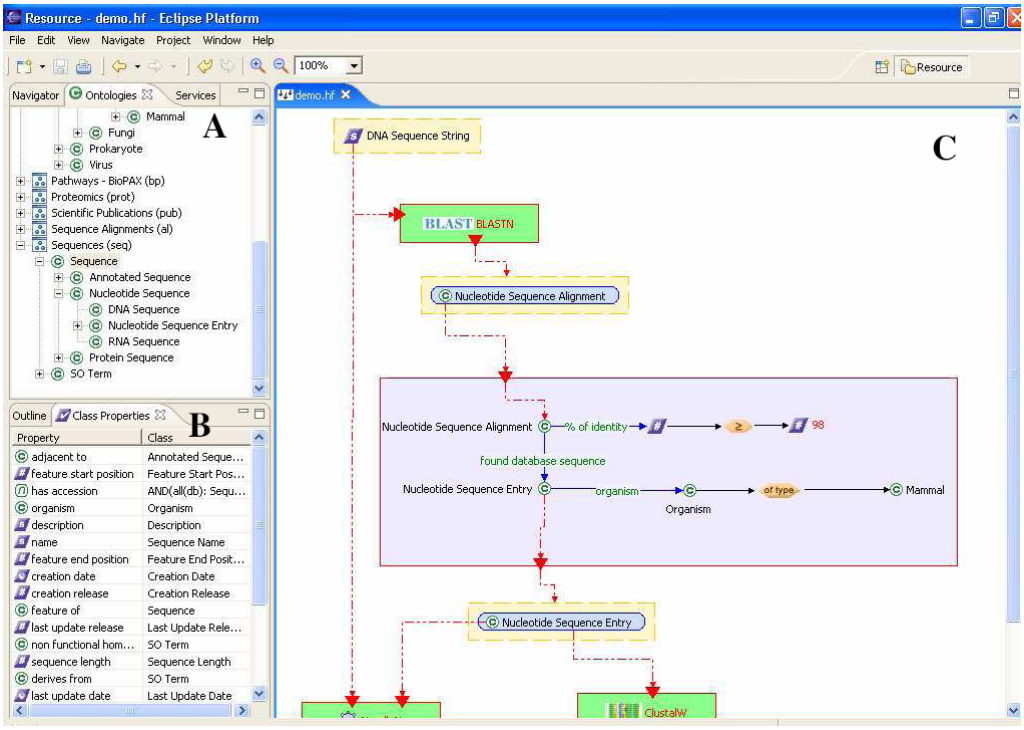


Figure 40: HyperFlow Framework. (a) Ontology from which the user can create the query. (b) Properties of the *Sequence* type selected in (a). (c) Workflow/query graph builder. The part that is really a query graph is highlighted by the blue rectangle. (Example from [DOT2006]).

Figure 41 presents a table comparing the expressivity power of various VQL-based systems. This could be of interest to determine the functionalities to implement for ACGT's VQL system.

	HyperFlow	QUIVER	Kaleidoquery	Gql	QGraph	MDDQL	HVQS	VOQL	VOODOO	VQE	QBE	DFQL	Iconic SQL
Data model	OO (OWL-based)	OO (ODMG)	OO (ODMG)	OO (functional)	OO (ODMG)	OO (ontology)	Relational	OO (ODMG)	OO (ODMG)	OO (ODMG)	Relational	Relational	Relational
Language Paradigm	Graph + Data Flow	Graph + Data Flow	Filter Flow	Graph	Graph + Text	Graph	Graph	Graph + Text	Frames	Frames	Frames	Data Flow	Iconic
Completely Visual	✓	✓	✓	✓		✓	✓		✓	✓			✓
Projection	✓	✓	✓	✓	✓	✓	✓	✓/NV	✓	✓	✓	NV	✓
Binary Constraints	✓	✓	✓	✓	✓	P		✓/NV	NV	✓	NV	NV	✓
IN (constants)	✓		✓	✓				NV			NV	NV	
Disjunction	✓		✓	✓	P	?	P	NV			✓	NV	✓
Negation	✓		✓	✓	✓	?		NV			✓	NV	✓
Relationships	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Arbitrary joins	✓	✓	✓	✓				✓/NV	✓	✓	NV		
Outer Joins	✓												
Existential Quantifications	✓		✓				✓	✓	✓			NV	
Universal Quantification	✓		✓				✓	✓	✓				
Transitive properties	✓												
Group By	✓		✓	✓	✓		✓		✓		✓		
Having	✓		✓		P				P		✓		
Aggregate Functions	✓	✓	✓	✓			✓		✓	✓	✓	✓	
Binary Set Operators	✓	✓	✓	✓	P		✓	✓			P	✓	
Collection Operators*	✓	✓											
Arithmetic Operators	✓	✓	✓	✓				NV		✓	NV	NV	
Custom Functions / Methods	✓	✓							NV			✓	
Order By	✓		✓						✓		✓	NV	✓
Subqueries	✓	✓	✓		P		✓		P			P	
Distinct	✓			✓			✓		✓		✓	NV	✓
OF TYPE	✓				✓	✓		✓					
Construct	✓	✓											
Construct Graph	✓												
Output Field Aliases	✓	✓											
Recursion	✓												
Closure	✓						✓	P		✓			

Figure 41. Comparison of VQL systems expressivity (from [DOT2006]). NV – supported in a non visual manner. P – Partially supported. * - Collection operators – listtoset, flatten, element, etc.

20.6 References

[SLO2003] P.M.A. Sloot, G.D. van Albada, E.V. Zudilova, P. Heinzlreiter, D. Kranzlmüller, H. Rosmanith, J. Volkert: Grid-based Interactive Visualisation of Medical Images. Proceedings of the First European HealthGrid Conference, pp. 57- 66 (2003).

[BRO2004a] K. Brodlie, D. Duce, J. Gallop, M. Sagar, J. Walton and J. Wood: Visualization in Grid Computing Environments. IEEE Visualization 2004, pp. 155–162 (2004).

[BRO2004b] K. Brodlie, J. Wood, D. Duce, M. Sagar: gViz - Visualization and Computational Steering on the Grid. Proceedings of the UK e-Science All Hands Meeting 2004, pp. 54-60. ISBN 1-904425-21-6 (2004).

[CAR1989] N. Carriero and D. Gelernter: Linda in Context. Communications of the ACM, 32(4), pp. 444-458 (1989).

[CHA2004] S.M. Charters, N.S. Holliman, M. Munro: Visualisation on the Grid - A Web Service Approach. In Proceedings of the UK e-Science All Hands Meeting 2004, ISBN 1-904425-21-6 (2004).

[CZA2004] K. Czajkowski, D. F. Ferguson, I. Foster, J. Frey, S. Graham, I. Sedukhin, D. Snelling, S. Tuecke and W. Vambenepe: The WS-Resource Framework (white paper), 2004. On the web: http://www.globus.org/wsrf/specs/ws-wsrf.pdf

[GEL1985] D. Gelernter: Generative Communication in Linda. ACM Trans. Program. Lang. Syst. 7(1): 80-112 (1985).

[SCA2006] M. Scarpa, R.G. Belleman, P.M.A. Sloot, C.T.A.M. de Laat: Highly Interactive Distributed Visualization. iGrid2005 special issue of Future Generation Computer Systems, (2006).

[SHA2003] J. Shalf and E.W. Bethel: The Grid and Future Visualization System Architectures. IEEE Computer Graphics and Applications, 23(2), pp. 6-9, March 2003.

[BUT2005] Butler G., Wang G., Wang Y. and Zou L. A graph database with visual queries for genomics. Proceedings of the 3rd Asia-Pacific Bioinformatics Conference (APBC2005): 17-21 January 2005, Singapore.

[CON1997] Consens M. and Mendelzon A. Hy+: a Hygraph-based query and visualization system. SIGMOD Rec 1997, 22(2), 511-516.

[DOT2005] Dotan, D. and Pinter, RY. HyperFlow: an Integrated Visual Query and Data-Flow Language for End-User Information Analysis. Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC'05), September 2005, pp. 27-34.

[DOT2006] Dotan, D. (2006). HyperFlow: a Visual, Ontology-Based Query and Data-Flow Language for End-User Information Analysis. Master of Science in Computer Science thesis. Technion - Israel Institute of Technology, Haifa, Israel.

[DUR2006] Durand P, Labarre L, Meil A, Divol JL, Vandenbrouck Y, Viari A, Wojcik J (2006) GenoLink: a graph-based querying and browsing system for investigating the function of genes and proteins. BMC Bioinformatics. 7(1):21

[ETZ1996] Etzold T, Ulyanov 4 Argos P (1996). SRS: Information retrieval system for molecular biology data banks. Methods Enzymol, 266:114-128.

[GEM1993] Gemis M., Paredaens J., Thyssens I. and Van den Bussche J. GOOD: a graph-oriented object database system. SIGMOD Rec 1993, 22(2), 505-510.

[GEN2000] Genoud, P., Dupierris, V., Page, M., Bruley, C., Ziebelin, D., Gensel, J. and Bardou, D. From AROM, a new object based knowledge representation system, to WebAROM, a knowledge bases server. 9th Int. Conf. on Artificial Intelligence: Methodology, Systems, and Applications: 20-23 September 2000, Varna, Bulgaria.

(http://www.inrialpes.fr/sherpa/arom/index.html)

[HAA2001] Haas LM, Schwarz PM, Kodali P, Kotlar E, Rice JE, Swope WC (2001). DiscoveryLink: A system for integrated access to life sciences data sources. IBM Syst J, 40:489-511.

[PAP1994] Papantonakis A. and King P.J.H Gql, a declarative graphical query language based on the functional data model. Proceedings of the workshop on Advanced visual interfaces (AVI '94): June 1-4, 1994, Bari, Italy, 113-122.

[POU2001] Poulovassilis A. and Hild S. Hyperlog: A Graph-Based System for Database Browsing, Querying, and Update. IEEE Trans. Knowl. Data Eng 2001, 13(2), 316-333.

[STE2000] Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, Goble CA, Bras A (2000). TAMBIS: Transparent access to multiple bioinformatics information sources. Bioinformatics, 16:184-185.

[WWWNOR] Snow. http://www.northbears.org/.

[ZLO1977] Zloof MM (1977). Query by example - a data base language. IBM Syst J, 16(4):324-343.

21 Ethico-legal issues

21.1 Legal and Ethical Issues in ACGT

ACGT aims to deliver to the cancer research community an integrated Clinico-Genomic ICT environment enabled by a powerful Grid infrastructure. The technological platform will be validated in concrete setting of advanced **clinical trials** on **cancer**. Hence pilot trials have been selected based on the presence of clear research objectives, raising the need to integrate data at all levels of the human being (molecular, tissue, organ, patient, disease, individuals, group of individuals). Since *ACGT* promotes the principle of open source and open access, thus enabling the gradual creation of a European Biomedical Grid on Cancer, the project plans to introduce additional clinical trials during its lifecycle.

The objective of ACGT is to obtain a better understanding of the optimal adjuvant therapy for the individual patient through translational research. In the area of adjuvant systemic therapy for cancer the three most important tasks can be defined as follows:

- assessment of risk for metastasis (*prognosis*);
- assignment of differential risk to different groups of patients (*patient stratification*);
- selection of treatment for the individual patient (*individualized therapy*).

Research pertinent to these tasks will be performed on tumour and blood samples collected from patients involved in pilot trials. Such biomaterials are, on the one hand, valuable resources for biomedical research. On the other hand, they are part of the donor's body. At least as long as such samples can be traced back to the donor, they are carrier of personal and sensitive information and therefore protected by personality rights. Furthermore, since research using these materials can in principle reveal information which may be of far reaching importance for the donor, he or she needs to be protected against the misuse of such information. Therefore collection and use of such samples in research is regulated on the international level and in many cases by national law.

One of the essential preconditions for establishing an integrated Clinico-Genomic ICT environment employing data extracted from human tissues therefore is that all research done in this context involving human subjects conforms to existing legal and ethical requirements. In addition, new ethical and legal challenges coming along with such an integrated Clinico-Genomic ICT-environment have to be identified and met with appropriate measures. For example, some of the genetic information proceeded within ACGT might not only be of relevance for the patient but also for his (possibly not yet born) inheritors.

This chapter first depicts patient's rights which could be affected in the context of ACGT and second reviews current European and national legislation as well as relevant international documents and instruments pertinent to the protection of patients in the context of clinical trials, biomedical and genomic research.

21.1.1 Protections of patients and patient's rights

Clinical trials and biomedical research involving human beings may pose risks yet unknown to patients enrolled in such trials. The ethical and legal issues do be dealt with in the context of ACGT therefore are essentially related to specific and fundamental rights of the person:

- Right to the integrity of the person;
- Right to self-determination (*privacy*);
- Right to informational self-determination (*data protection*)

In this section, issues related to these rights will shortly be discussed.

21.1.2 Integrity of the person

Physical and mental integrity of a person is the most fundamental right a person can claim for himself/herself. One of the most important issues to be dealt with in the context of ACGT therefore is whether there is a threat for the physical integrity of patients involved in the pilot trials. These trials constitute the setting in which the technological platform to be developed in ACGT will be validated.

Currently, four pilot trials participants are included in ACGT: three on breast cancer (University of Oxford, Oxford [UK], Institut Jules Bordet, Brussels [Belgium], University of Crete, Heraklion [Greece]) and one on childhood nephroblastoma (University of Saarland, Homburg [Germany]). These trials involve – among other goals – the examination of new drugs and/or treatment regimes for cancer.

These clinical trials as such are not part of ACGT but have been set up outside and independent of this project. They are governed by national law and have been reviewed by local ethics committees or institutional review boards. Practices and procedures involved in these clinical trials are subjected to the overview of the relevant national bodies and do not fall directly into the responsibility of ACGT. Furthermore, all groups which conduct clinical trials and contribute biomaterials and data to ACGT have declared to adhere to relevant guidelines and to take appropriate measures to prevent or minimize possible harm for the patients.

In the context of ACGT, patients who have agreed to participate in these clinical trials and have given valid consent to it are asked to donate blood and/or tumour tissue for additional research. For the recovery of such samples standard medical procedures are used:

- *Tumours* usually are surgically removed during (experimental or regular) treatment and subjected to diagnostic examination. Part of the tumour tissue will than be used for genomic analysis carried out for the purposes of ACGT. Therefore, extraction of tumour tissue for ACGT does not involve risks beyond the one posed by the treatment which is carried out for the benefit of the patient and to which the patient has agreed to.

- *Blood samples* may be collected for clinico-genomic research in addition to the samples which have been extracted for diagnostic purposes. This usually does not involve more than minimal risk to the patient.

Bearing these procedures in mind, clinico-genomic research which is specifically conducted for the purposes of ACGT should not pose more than minimal risk to the physical integrity of the person, if carried out according to the provisions prescribed by law (see chapter 2.3).

21.1.3 Self determination and informational self-determination

The second fundamental right which is at stake in the context of biomedical and clinico-genomic research is the right to self-determination and – associated with it – the right to informational self-determination. Both acknowledge that human beings are autonomous

persons which have the right to decide for themselves. Decisions to take part in clinical trials involving new drugs or treatment regimes, or to donate body material for research can – in principle – have far reaching consequences for the individual and his or her family. Therefore, such a decision must be self-determined and autonomous.

Two conditions have to be fulfilled in order to realize autonomy: *agency* and *liberty*.

- *Agency*: the individual has to be able to understand what is asked of him/her and to act intentional – it must possess agency. Sometimes people with cognitive impairments lack this ability to understand and to act intentionally. This poses specific problems if research is to be conducted with children or mentally incapacitated individuals. In these cases, international and national law provides specific provisions which have to be observed when research is carried out with minors.

- *Liberty*: the decision to take part in clinical research must be voluntary and free. It must not be subjected to social or moral pressure or to undue economic incentives.

One essential precondition which must be fulfilled in order to exercise autonomy in the context of clinical or clinico-genomic research is that the patient must be aware of what he or she is agreeing to. Therefore, before consent is asked from patients, they have to be adequately informed about the nature of the trial, possible risks and benefits, treatment alternatives, their right to withdraw from the trial and other rights which are specified by international and/or national guidelines.

In general, information has to be

- *of high quality*: up to date, interdisciplinary

- *appropriate:* with respect to scope and comprehensiveness, and

- *neutral*: balanced with respect to the presentation of chances and risks.

As a consequence, when subjects enrolled in clinical trials are asked in addition to donate blood or tumour tissue for genetic or genomic research they should be adequately informed about the type of research. This may be especially relevant if the samples are planned to be stored for a long time, and used for different research projects.

Up to now, there is no consensus in the international ethical and legal discourse how specific this information has to be, whether “biomedical research” is sufficiently precise as a purpose, or whether more details on the prospective research have to be given. ACGT will deal with this and other upcoming questions elsewhere.

21.2 Review of current law, guidelines and documents

Efficient protection of patients and their rights in all stages of ACGT requires strict adherence to the requirements imposed by existing European and National Legislation is of paramount importance.

To be able to ensure the full compliance of ACGT with all relevant legal and ethical issues, in-depth knowledge of the existing legislation, both at national, European and international levels gas been acquired as a first step.

In the sections to follow, European legislation, international documents and national law pertinent to ACGT is depicted and relevant paragraphs and articles are briefly described.

21.2.1 European Legislation

21.2.1.1 CHARTER OF FUNDAMENTAL RIGHTS OF THE EUROPEAN UNION

The Charter of Fundamental Rights of the European Union was solemnly proclaimed by the European Council in 2000 and was also approved by the European Commission and the European Parliament. It is part of the proposed European Constitution that failed to be ratified. Therefore the Charter contains non-binding law. But nevertheless it is an important guideline of interpretation. The most important provisions concerning ACGT are:

Article 3: Right to the integrity of the person

Article 3, which refers to the right to the integrity of the person, states that:

1. Everyone has the right to respect for his or her physical and mental integrity.

2. In the fields of medicine and biology, the following must be respected in particular:
   - the free and informed consent of the person concerned, according to the procedures laid down by law,
   - the prohibition of eugenic practices, in particular those aiming at the selection of persons,
   - the prohibition on making the human body and its parts as such a source of financial gain,
   - the prohibition of the reproductive cloning of human beings.

Article 7: Respect for private and family life

The Charter of Fundamental Rights of the European Union protects the right to respect for private life. It echoes in some extent the right to self-determination and to the right to informational self-determination. This legal tool states precisely that:

> *Everyone has the right to respect for his or her private and family life, home and communications.*

Article 8: Protection of personal data

Also in Article 8, related to the protection of personal data, it is stated that:

1. Everyone has the right to the protection of personal data concerning him or her.

2. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified.

3. Compliance with these rules shall be subject to control by an independent authority.

21.2.1.2 DIRECTIVE 95/46/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL OF 24 OCTOBER 1995

The Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free

movement of such data created a legal framework common to all the Member States relative to the processing of personal data. This Directive had to be transposed into national law by the Member States[16].

In a first step, the Directive lays down the rules applicable to every processing of personal data. In a second step, the Directive provides some additional rules for the processing of sensitive data (as medical data). In a third step, the Directive provides special rights of the data subject and control mechanisms. In a fourth and last step, the Directive rules the transfer of personal data to third countries.

The Directive provides the general rules on the lawfulness of the processing of personal data. This covers:

- the principles relating to data quality (art. 6),
- the criteria for making data processing legitimate (art. 7),
- the special categories of processing (art. 8-9),
- the information to be given to the data subject (art. 10-11),
- the data subject's right of access to data (art. 12) (cf. art. 13 for exemptions and restrictions),
- the data subject's right to object (art. 14-15),
- the confidentiality and the security of processing (art. 16-17),
- the notification to the supervisory authority (art. 18-21),

Then the Directive covers the issues of judicial remedies, liability and sanctions (art. 22-24).

Transfer of personal data to third countries is subject to special rules (art. 25-26).

The Directive also encourages the drafting of Codes of conducts (art. 27).

Finally the Directive creates supervisory authorities and the working party on the protection of individuals with regard to the processing of personal data (art. 28-30).

Some rules concern more directly the ACGT Project and need to be stressed:

- With special reference to medical research it is clearly stated that the prohibition of processing of sensitive personal data of Article 8(1) may be lifted for reasons of substantial public interest, by national law or decision of the supervisory authority if Member States provide suitable safeguards (Article 8(4));

- Also for medical research the prohibition of processing of sensitive personal data may be lifted, if this medical research could be considered in some special cases to be a subcategory of preventive medicine, medical diagnosis, the provision of care or treatment, or management of health-care services, provided that the processing

[16] See also for the Status of implementation of Directive 95/46 on the Protection of Individuals with regard to the Processing of Personal Data in the different Member States: http://ec.europa.eu/justice_home/fsj/privacy/law/implementation_en.htm and for the transposition of the Directive in the Member States: http://ec.europa.eu/justice_home/fsj/privacy/lawreport/index_en.htm.

of sensitive personal data involved is carried out by a health professional or another person subject to an equivalent "obligation of secrecy" per national law or rules established by national competent bodies (Article 8(3)).

- Another specific tool for medical research is set by the Codes of conduct: Article 27(1) requires Member States and the Commission to encourage the drawing up of codes of conduct to assist with the implementation of the Directive in specific sectors of processing, representing categories of data controllers and to consult with data subjects or their representatives (Article 27(2)). Article 27(3) provides a role for the Article 29 Working Party in approving draft Community Codes and amendments to existing Community codes.

- Member States have to determine the processing operations likely to present specific risks to the rights and freedoms of data subjects and shall check that these processing operations are examined prior to the start thereof (art. 20.1).

- Such prior checks have to be carried out by the supervisory authority following receipt of a notification from the controller or by the data protection official, who, in cases of doubt, must consult the supervisory authority (art. 20.2).

  Member States may also carry out such checks in the context of preparation either of a measure of the national parliament or of a measure based on such a legislative measure, which define the nature of the processing and lay down appropriate safeguards (art. 20.3).

For the purposes of this Directive:

(a) '***personal data***' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity;

(b) '***processing of personal data***' ('processing') shall mean any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction;

It is very stated very clearly in the directive (Section 1, Article 6) that

- personal data must be:

  1. processed fairly (in compliance with the announced purposes of the data processing) and lawfully (this latter referring in ACGT notably to the respect of the medical secrecy);

  2. collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes (this refers to the proportionality test). Further processing of data for historical, statistical or scientific purposes shall not be considered as incompatible provided that Member States provide appropriate safeguards;

3. adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed (this refers also to the proportionality test);

4. kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed (right to oblivion). Member States shall lay down appropriate safeguards for personal data stored for longer periods for historical, statistical or scientific use.

- Personal data may only be processed under the conditions described in article 7 for the processing of "simple" personal data. The unambiguously consent of the data subject is the first condition allowing the processing of "simple" personal data.

- The processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex life, is banned (article 8). This prohibition may be lifted under the conditions described in article 8.2. The explicit consent of the data subject is the first condition allowing the processing of medical data (art. 8.2.a). But Member State may provide that this prohibition may not be lifted by the data subject's consent.

- Processing of medical data can be legalised by the Member States, if the data is required for the purposes of preventive medicine, medical diagnosis, the provision of care or treatment or the management of health-care services, and where those data are processed by a health professional subject under national law or rules established by national competent bodies to the obligation of professional secrecy or by another person also subject to an equivalent obligation of secrecy.

The Directive makes additional significant provisions, especially about the data subject's right of information and its right of access to the data, as described below:

Section 4, Article 10: Information in cases of collection of data from the data subject

Member States shall provide that the controller or his representative must provide a data subject from whom data related to him/herself are collected with at least the following information, except where he already has it:

(a) the identity of the controller and of his representative, if any;

(b) the purposes of the processing for which the data are intended;

(c) any further information such as - the recipients or categories of recipients of the data, - whether replies to the questions are obligatory or voluntary, as well as the possible consequences of failure to reply, - the existence of the right of access to and the right to rectify the data concerning him o in so far as such further information is necessary, having regard to the specific circumstances in which the data are collected, to guarantee fair processing in respect of the data subject.

Section 4, Article 11: Information where the data have not been obtained from the data subject

The Directive makes the following provisions where the data have not been obtained from the data subject,

1. Member States shall provide that the controller or his representative must at the time of undertaking the recording of personal data or if a disclosure to a third party is envisaged, no later than the time when the data are first disclosed provide the data subject with at least the following information, except where he already has it:

    (a) the identity of the controller and of his representative, if any;

    (b) the purposes of the processing;

    (c) any further information such as - the categories of data concerned, - the recipients or categories of recipients, - the existence of the right of access to and the right to rectify the data concerning him in so far as such further information is necessary, having regard to the specific circumstances in which the data are processed, to guarantee fair processing in respect of the data subject.

2. Paragraph 1 shall not apply where, in particular for processing for statistical purposes or for the purposes of historical or scientific research, the provision of such information proves impossible or would involve a disproportionate effort or if recording or disclosure is expressly laid down by law. In these cases Member States shall provide appropriate safeguards.

Section 5, Article 12

Member States shall guarantee every data subject the right to obtain from the controller without constraint at reasonable intervals and without excessive delay or expense:

(a) confirmation as to whether or not data relating to him are being processed and information at least as to the purposes of the processing, the categories of data concerned, and the recipients or categories of recipients to whom the data are disclosed,

(b) communication to him in an intelligible form of the data undergoing processing and of any available information as to their source,

(c) knowledge of the logic involved in any automatic processing of data concerning him at least in the case of the automated decisions referred to in Article 15 (1);

The confidentiality and the security of the data processing must be guaranteed. There are special rules when the processing is carried out by a processor on behalf of the data controller.

Chapter 4, Article 25: Transfer of personal data to third countries.

The Member States shall provide that the transfer to a third country of personal data which are undergoing processing or are intended for processing after transfer may take place only if, without prejudice to compliance with the national provisions adopted pursuant to the other provisions of this Directive, the third country in question ensures an adequate level of protection.

The adequacy of the level of protection afforded by a third country shall be assessed in the light of all the circumstances surrounding a data transfer operation or set of data transfer operations; particular consideration shall be given to the nature of the data, the purpose and duration of the proposed processing operation or operations, the country of origin and country of final destination, the rules of law, both general and sectoral, in force in the third country in question and the professional rules and security measures which are complied with in that country.

The Member States and the Commission shall inform each other of cases where they consider that a third country does not ensure an adequate level of protection within the meaning of paragraph 2.

Where the Commission finds, under the procedure provided for in Article 31 (2), that a third country does not ensure an adequate level of protection within the meaning of paragraph 2 of this Article, Member States shall take the measures necessary to prevent any transfer of data of the same type to the third country in question.

Article 26: Derogations

1. By way of derogation from Article 25 and save where otherwise provided by domestic law governing particular cases, Member States shall provide that a transfer or a set of transfers of personal data to a third country which does not ensure an adequate level of protection within the meaning of Article 25 (2) may take place on condition that:

    a. the data subject has given his consent unambiguously to the proposed transfer; or

    b. the transfer is necessary for the performance of a contract between the data subject and the controller or the implementation of pre-contractual measures taken in response to the data subject's request; or

    c. the transfer is necessary for the conclusion or performance of a contract concluded in the interest of the data subject between the controller and a third party; or

2. Without prejudice to paragraph 1, a Member State may authorize a transfer or a set of transfers of personal data to a third country which does not ensure an adequate level of protection within the meaning of Article 25 (2), where the controller adduces adequate safeguards with respect to the protection of the privacy and fundamental rights and freedoms of individuals and as regards the exercise of the corresponding rights; such safeguards may in particular result from appropriate contractual clauses.

3. The Member State shall inform the Commission and the other Member States of the authorizations it grants pursuant to paragraph 2.

21.2.1.3 DIRECTIVE 2001/20/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 4 April 2001

Directive 2001/20/EC of the European Parliament and of the Council of 4 April 2001 on the approximation of the laws, regulations and administrative provisions of the Member States relating to the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use.

This Directive establishes specific provisions regarding the conduct of clinical trials, including multi-centre trials, on human subjects involving medicinal products and is the most important European directive with respect to clinical trials.

The directive provides useful definitions and provisions for clinical trials involving minors:

Article 2: Definitions

(a) 'informed consent': decision, which must be written, dated and signed, to take part in a clinical trial, taken freely after being duly informed of its nature, significance, implications and risks and appropriately documented, by any person capable of giving consent or, where the person is not capable of giving consent, by his or her legal representative; if the person concerned is unable to write, oral consent in the presence of at least one witness may be given in exceptional cases, as provided for in national legislation.

Article 4: Clinical trials on minors

In addition to any other relevant restriction, a clinical trial on minors may be undertaken only if:

(a) the informed consent of the parents or legal representative has been obtained; consent must represent the minor's presumed will and may be revoked at any time, without detriment to the minor;

(b) the minor has received information according to its capacity of understanding, from staff with experience with minors, regarding the trial, the risks and the benefits;

(c) the explicit wish of a minor who is capable of forming an opinion and assessing this information to refuse participation or to be withdrawn from the clinical trial at any time is considered by the investigator or where appropriate the principal investigator;

(d) no incentives or financial inducements are given except compensation;

(e) some direct benefit for the group of patients is obtained from the clinical trial and only where such research is essential to validate data obtained in clinical trials on persons able to give informed consent or by other research methods; additionally, such research should either relate directly to a clinical condition from which the minor concerned suffers or be of such a nature that it can only be carried out on minors;

(f) the corresponding scientific guidelines of the Agency have been followed;

(g) clinical trials have been designed to minimise pain, discomfort, fear and any other foreseeable risk in relation to the disease and developmental stage; both the risk threshold and the degree of distress have to be specially defined and constantly monitored;

(h) the Ethics Committee, with paediatric expertise or after taking advice in clinical, ethical and psychosocial problems in the field of paediatrics, has endorsed the protocol; and

(i) the interests of the patient always prevail over those of science and society.

21.2.1.4 DIRECTIVE 98/79/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 October 1998 on in vitro diagnostic medical devices

Article 1: Scope, definitions

For the purposes of this Directive, the following definitions shall apply:

(a) 'medical device' means any instrument, apparatus, appliance, material or other article, whether used alone or in combination, including the software necessary for its proper application, intended by the manufacturer to be used for human beings for the purpose of:

- diagnosis, prevention, monitoring, treatment or alleviation of disease,
- diagnosis, monitoring, treatment, alleviation or compensation for an injury or handicap,
- investigation, replacement or modification of the anatomy or of a physiological process,
- control of conception,

and which does not achieve its principal intended action in or on the human body by pharmacological, immunological or metabolic means, but which may be assisted in its function by such means;

(b) 'in vitro diagnostic medical device' means any medical device which is a reagent, reagent product, calibrator, control material, kit, instrument, apparatus, equipment, or system, whether used alone or in combination, intended by the manufacturer to be used in vitro for the examination of specimens, including blood and tissue donations, derived from the human body, solely or principally for the purpose of providing information:

- concerning a physiological or pathological state, or
- concerning a congenital abnormality, or
- to determine the safety and compatibility with potential recipients, or
- to monitor therapeutic measures.

21.2.1.5 CONVENTION OF THE COUNCIL NO. 05 FOR THE PROTECTION OF HUMAN RIGHTS AND FUNDAMENTAL FREEDOMS

This Convention was not ratified by the European Union itself, but by all of the Member States. According to Art. 6 para. 2 of the Treaty on European Union the Union shall respect the Convention as general principles of Community Law. The most important provision concerning ACGT is:

Article 8 – Right to respect for private and family life

1. Everyone has the right to respect for his private and family life, his home and his correspondence.

2. There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.

21.2.1.6 CONVENTION NO. 108 OF THE COUNCIL OF EUROPE FOR THE PROTECTION OF INDIVIDUALS WITH REGARD TO AUTOMATIC PROCESSING OF PERSONAL DATA

The Convention was not ratified by the European Union itself, but by all Member States. It obliges the signing Member States to transpose the provisions into national law. The most important provisions concerning ACGT are:

Article 1 – Object and purpose

The purpose of this convention is to secure in the territory of each Party for every individual, whatever his nationality or residence, respect for his rights and fundamental freedoms, and in particular his right to privacy, with regard to automatic processing of personal data relating to him ("data protection").

Article 2 – Definitions

For the purposes of this convention:

(a) "personal data" means any information relating to an identified or identifiable individual ("data subject");

(b) "automated data file" means any set of data undergoing automatic processing;

(c) "automatic processing" includes the following operations if carried out in whole or in part by automated means: storage of data, carrying out of logical and/or arithmetical operations on those data, their alteration, erasure, retrieval or dissemination;

(d) "controller of the file" means the natural or legal person, public authority, agency or any other body who is competent according to the national law to decide what should be the purpose of the automated data file, which categories of personal data should be stored and which operations should be applied to them.

Article 4 – Duties of the Parties

1. Each Party shall take the necessary measures in its domestic law to give effect to the basic principles for data protection set out in this chapter.

2. These measures shall be taken at the latest at the time of entry into force of this convention in respect of that Party.

Article 5 – Quality of data

Personal data undergoing automatic processing shall be:

(a) obtained and processed fairly and lawfully;

(b) stored for specified and legitimate purposes and not used in a way incompatible with those purposes;

(c) adequate, relevant and not excessive in relation to the purposes for which they are stored;

(d) accurate and, where necessary, kept up to date;

(e) preserved in a form which permits identification of the data subjects for no longer than is required for the purpose for which those data are stored.

Article 6 – Special categories of data

Personal data revealing racial origin, political opinions or religious or other beliefs, as well as personal data concerning health or sexual life, may not be processed automatically unless domestic law provides appropriate safeguards. The same shall apply to personal data relating to criminal convictions.

Article 7 – Data security

Appropriate security measures shall be taken for the protection of personal data stored in automated data files against accidental or unauthorised destruction or accidental loss as well as against unauthorised access, alteration or dissemination.

Article 9 – Exceptions and restrictions

1. No exception to the provisions of Articles 5, 6 and 8 of this convention shall be allowed except within the limits defined in this article.

2. Derogation from the provisions of Articles 5, 6 and 8 of this convention shall be allowed when such derogation is provided for by the law of the Party and constitutes a necessary measure in a democratic society in the interests of:

    (a) protecting State security, public safety, the monetary interests of the State or the suppression of criminal offences;

    (b) protecting the data subject or the rights and freedoms of others.

3. Restrictions on the exercise of the rights specified in Article 8, paragraphs b, c and d, may be provided by law with respect to automated personal data files used for statistics or for scientific research purposes when there is obviously no risk of an infringement of the privacy of the data subjects.

Article 10 – Sanctions and remedies

Each Party undertakes to establish appropriate sanctions and remedies for violations of provisions of domestic law giving effect to the basic principles for data protection set out in this chapter.

Article 11 – Extended protection

None of the provisions of this chapter shall be interpreted as limiting or otherwise affecting the possibility for a Party to grant data subjects a wider measure of protection than that stipulated in this convention.

Article 12 – Transborder flows of personal data and domestic law

1. The following provisions shall apply to the transfer across national borders, by whatever medium, of personal data undergoing automatic processing or collected with a view to their being automatically processed.

2. A Party shall not, for the sole purpose of the protection of privacy, prohibit or subject to special authorisation transborder flows of personal data going to the territory of another Party.

3. Nevertheless, each Party shall be entitled to derogate from the provisions of paragraph 2:

   (a) insofar as its legislation includes specific regulations for certain categories of personal data or of automated personal data files, because of the nature of those data or those files, except where the regulations of the other Party provide an equivalent protection;

   (b) when the transfer is made from its territory to the territory of a non EU belonging State through the intermediary of the territory of another Party, in order to avoid such transfers resulting in circumvention of the legislation of the Party referred to at the beginning of this paragraph.

21.2.1.7 CONVENTION No. 164 OF THE COUNCIL OF EUROPE FOR THE PROTECTION OF HUMAN RIGHTS AND DIGNITY OF THE HUMAN BEING WITH REGARD TO THE APPLICATION OF BIOLOGY AND MEDICINE (Convention on Human Rights and Biomedicine)

Relevant Text Excerpts:

Chapter 2, Article 5 - General rule

An intervention in the health field may only be carried out after the person concerned has given free and informed consent to it. This person shall beforehand be given appropriate information as to the purpose and nature of the intervention as well as on its consequences and risks. The person concerned may freely withdraw consent at any time.

Chapter 2, Article 6 - Protection of persons not able to consent

Subject to Articles 17 and 20 below, an intervention may only be carried out on a person who does not have the capacity to consent, for his or her direct benefit.

Where, according to law, a minor does not have the capacity to consent to an intervention, the intervention may only be carried out with the authorisation of his or her representative or an authority or a person or body provided for by law.

The opinion of the minor shall be taken into consideration as an increasingly determining factor in proportion to his or her age and degree of maturity.

The representative, the authority, the person or the body mentioned in paragraphs 2 and 3 above shall be given, under the same conditions, the information referred to in Article 5.

Chapter 4, Article 12 - Predictive genetic tests

Tests which are predictive of genetic diseases or which serve either to identify the subject as a carrier of a gene responsible for a disease or to detect a genetic predisposition or susceptibility to a disease may be performed only for health purposes or for scientific research linked to health purposes, and subject to appropriate genetic counselling.

21.2.1.8 Recommendations

Apart from the EC Directives above, ACGT will also take into account also the following provisions:

- ***Council of Europe, Recommendation No. R(97)5 on the protection of medical data adopted of 13 February 1997.***

  ACGT will be strictly compliant to the provisions of article 4, in particular Medical data will be collected and processed (…) for preventive medical purposes or for diagnostic or for therapeutic purposes.

- ***Council of Europe, Recommendation on human rights and biomedicine, concerning biomedical research, Strasbourg 25th of January 2005.***

  Additional protocol to the convention on human rights and biomedicine, concerning biomedical research, Strasbourg 25th of January 2005 of the Council of Europe (CETS No 195) covering the full range of research activities in the health field involving interventions on human being and in particular the primacy of the human being (Chapter II article 3), Chapter III ethics committee , Chapter IV Information and consent, Chapter V (protection of persons not able to consent to research), Chapter VII (Safety and Supervision) Chapter VIII (Confidentiality and right to information).

21.2.2 Relevant International Instruments and Documents

At the international level, a number of documents and instruments exist which have been issued by professional bodies or international organisations. Although they are not legally binding they are nevertheless important, since in most cases they are the result of a consensus process involving numerous individuals and groups concerned with the rights and the wellbeing of patients taking part in clinical trials and/or biomedical research.

21.2.2.1 WORLD MEDICAL ASSOCIATION DECLARATION OF HELSINKI

The World Medical Association Declaration of Helsinki (Ethical Principles for Medical Research Involving Human Subjects) adopted by the 18th WMA General Assembly, Helsinki, Finland, June 1964, as amended by various Assemblies, last in Note of Clarification on Paragraph 30 added by the WMA General Assembly, Tokyo 2004.

In particular, part B (The Basic Principles for All Medical Research) contains some fundamental guidelines in terms of medical research and data treatment:

- Section 10: "It is the duty of the physician in medical research to protect the life, health, privacy, and dignity of the human subject."

- Section 21: "The right of research subjects to safeguard their integrity must always be respected. Every precaution should be taken to respect the privacy of the subject, the confidentiality of the patient's information and to minimize the impact of the study on the subject's physical and mental integrity and on the personality of the subject."

- Section 25: "When a subject deemed legally incompetent, such as a minor child, is able to give assent to decisions about participation in research, the investigator

must obtain that assent in addition to the consent of the legally authorized representative".

21.2.2.2 UNESCO DECLARATIONS

Important relevant documents and legislations have been also produced by UNESCO's Internationals Bioethics Committee (IBC) and adopted by the General Conference of UNESCO. Three of them are especially important in the context of ACGT and will be considered:

- the Universal Declaration on the Human Genome and Human Rights (1997) http://portal.unesco.org/en/ev.php-URL_ID=13177&URL_DO=DO_TOPIC&URL_SECTION=201.html,
- and the International Declaration of Human Genetic Data (2003) http://portal.unesco.org/shs/en/file_download.php/6016a4bea4c293a23e913de638045ea9Declaration_en.pdf, and
- Universal Declaration on Bioethics and Human Rights (2005). http://portal.unesco.org/shs/en/file_download.php/46133e1f4691e4c6e57566763d474a4dBioethicsDeclaration_EN.pdf

Especially the International Declaration on the Human Genome and Human Rights contains many Articles (3 through 23), which cover issues raised in the context of ACGT. Although declarations of UNESCO are legally not binding (similar to the World Medical Association Declaration of Helsinki), they have been accepted by the member states of UNESCO and therefore are relevant for most European countries as well.

21.2.2.3 ARTICLE 29 DATA PROTECTION WORKING PARTY. WORKING DOCUMENT ON GENETIC DATA. Adopted on March 17, 2004.

The Art. 29 Data Protection Working Party was established by Art. 29 of the Data Protection Directive 95/46/EC and is an independent advisory body.[17] It can make recommendations on all matters relating to the protection of persons on it's own initiative, advice the Commission on any amendment or specific measure to safeguard the rights and freedoms of natural persons with regard to the processing of personal data and on any other proposed Community measures affecting such rights and freedoms and give the Commission an opinion on the level of protection in the Community and in third countries according to Art. 30 of the Directive 95/46/EC and Art. 14 of Directive 97/66/EC.

This Working Paper contains the authoritative interpretation of the DIRECTIVE 95/46/EC with respect to genetic data and the purposes for which the collection and processing of such data may take place.

Relevant text excerpts of the Working Document on Genetic Data of the Art. 29 Data Protection Working Party:

Section II. DEFINITIONS AND MAIN CHARACTERISTICS OF GENETIC DATA

Definitions:

[17] See also the Website of Art. 29 Date Protection Working Party: http://ec.europa.eu/justice_home/fsj/privacy/workinggroup/index_en.htm.

- All data of whatever type concerning the hereditary characteristics of an individual or concerning the pattern of inheritance of such characteristics within a related group of individuals (Council of Europe Recommendation N°R(97)5)

- Any data concerning the hereditary characteristics of an individual or group of related individuals (Art 2 (g) of the 2 August 2002 law of Luxembourg on the protection of persons with regard to the processing of personal data)

- Non-obvious information about heritable characteristics of individuals obtained by analysis of nucleic acids or by other scientific analysis (International Declaration on Human Genetic data, UNESCO)

Genetic data thus present a number of characteristics which can be summarised as follows:

- while genetic information is unique and distinguishes an individual from other individuals, it may also at the same time reveal information about and have implications for that individual's blood relatives (biological family) including those in succeeding and preceding generations, Furthermore, genetic data can characterise a group of persons (e.g. ethnic communities);

- genetic data can reveal parentage and family links;

- genetic information is often unknown to the bearer him/herself and does not depend on the bearer's individual will since genetic data are non modifiable;

- genetic data can be easily obtained or be extracted from raw material although this data may at times be of dubious quality;

- taking into account the developments in research, genetic data may reveal more information in the future

- and be used by an ever increasing number of agencies for various purposes.

The Working Party also discusses whether genetic data are “personal data” and “sensitive data”:

Section III. APPLICABILITY OF THE 95/46/EC DIRECTIVE

According to Art 2 (a) of the Directive: "personal data" shall mean any information relating to an identifiable natural person (data subject); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity."

There is no doubt that genetic information content is covered by this definition. Indeed, a link to a specific person, i.e. the fact that the person concerned is identified or identifiable, is clear in the majority of cases. Nevertheless in some cases it is less clear, e.g. samples of DNA taken in a given place, such as traces at the scene of a crime. However, such samples may constitute a source of personal data in so far as it may be possible to associate samples of DNA with a given person, in particular once their origin has been confirmed by a court upon the forensic evidence. Therefore, in regulating genetic data, consideration should also be given to the legal status of DNA samples.

According to Article 8(1) of the Directive, categories of personal data whose sensitivity requires a higher level of protection includes "data concerning health". Genetic data may provide to an extent a detailed picture of a person's physical disposition and health condition and therefore could be considered as "data concerning health". Furthermore, genetic data may also describe specific forms of a wide range of physical characteristics. Thus, genetic data which determine the colour of someone's hair, for example, may not be regarded as data directly concerning health. In this context, genetic data can contribute e.g. to assess the ethnic origin of an individual and should as well be considered as falling within the scope of Art 8 (1).

Considering the extremely singular characteristics of genetic data and their link to information that may reveal the health condition or the ethnic origin, they should be treated as particularly sensitive data within the meaning of Article 8 (1) of the Directive and therefore be subject to the reinforced protection provided for in the Directive and the national laws transposing it.

Section IV: PURPOSES FOR WHICH THE COLLECTION AND PROCESSING OF GENETIC DATA MAY TAKE PLACE AND RELEVANT ISSUES

Due to the special nature and characteristics of genetic data and the impact their use may have on the individual's life and on the members of his family, it is very important to determine the purposes for which genetic data may be processed.

- Health care/ medical treatment
- …..
- Medical and scientific research.

Section V: CONCLUDING REMARKS

In Member States where the purposes and the appropriate safeguards for the processing of genetic data are not established by law, the data protection authorities (DPAs) are encouraged to play an even more active role in ensuring that the finality and proportionality principles of the Directive are fully respected.

In this respect, the Working Party recommends that Member States should consider submitting the processing of genetic data to prior checking by DPAs, in accordance with Article 20 of the Directive. This should in particular be the case with regard to the setting up and use of bio banks.

Moreover, closer cooperation and exchange of best practices between DPAs could prove to be an efficient way to compensate the present absence of regulatory framework in the field of the on-line "genetic testing direct to the public".

21.2.2.4 ARTICLE 29 DATA PROTECTION WORKING PARTY Opinion 6/2000 on the Human Genome and Privacy

The decoding of the DNA blueprint paves the way to new discoveries and uses in the field of genetic testing. On the other hand, the information can identify individuals, link them to others, and reveal complex data about the future health and development of those individuals and other people to whom they are genetically related. The Working Party wishes to emphasise the importance of privacy as a fundamental right and the consequent necessity of deploying new genetic technologies with safeguards adequate to protect that right.

21.2.2.5 OPINION OF THE EUROPEAN GROUP ON ETHICS IN SCIENCE AND NEW TECHNOLOGIES TO THE EUROPEAN COMMISSION, No. 11, 21 July 1998

Ethical Aspects of Human Tissue Banking [Relevant Text Excerpts]

Main ethical issues

Wherever tissues are removed from human beings, and possibly transplanted into other human beings, the activities involved in the collection and use of such tissues are subject to ethical requirements intended to safeguard respect for human beings, their dignity and autonomy, and for the common good.

21.2.2.6 INTERNATIONAL GUIDELINES FOR BIOMEDICAL RESEARCH INVOLVING HUMAN SUBJECTS.

Prepared by the Council for International Organizations of Medical Sciences (CIOMS) in collaboration with the World Health Organization (WHO). CIOMS, Geneva, 2002.

According with these Guidelines, the full respect of three basic ethical principles must be guaranteed: namely justice, respect for persons, and beneficence (maximizing benefits and minimizing harms and wrongs) or non-malevolence (doing no harm).

21.2.3 National Laws and Regulations

In the following section the national and international regulations that, in the Consortium's opinion, are relevant for the research are listed and the compliance of the project with the regulation explicitly stated.

ACGT involves pilots in Belgium, Germany, Greece and the UK as shown in the following figure.

The relevant laws and regulations in the countries where the research will be carried out are listed below.

UK

Data Protection Act 1998 Chapter 29. It contains all the regulations and, as far as research is concerned, Part IV specifies the Exemptions: 33 (4) Personal data which are processed only for research purposes are exempt from section 7 if:

(a) they are processed in compliance with the relevant conditions, and

(b) the results of the research or any resulting statistics are not made available in a form which identifies data subjects or any of them.

In the UK there is a national ethical approval application process as governed by COREC (Central Office for Research Ethics Committees). Individual Ethics Committees in the UK use the COREC processes to administer Ethics applications. ACGT will seek approvals from the Local Research Ethics Committee (LREC). http://www.corec.org.uk/applicants/apply/apply.htm.

With regard to informed consent the situation in the UK, and Oxford University in specifically, particularly relating to the samples we want to use in ACGT is as follows.

In Oxford hospital, using the nationally approved consent forms for surgery, there is a section specifically asking for excess tissue to be donated for research. Tissue samples for which a patient has declined consent are specifically earmarked. These will not be used.

With regard to analysis of agreed samples, they can only be used when the COREC approval has been obtained. This is a national scheme of research ethics committees and submission is to your local ethics committee which is monitored by national standards. The ethics committees function according to the new European regulations.

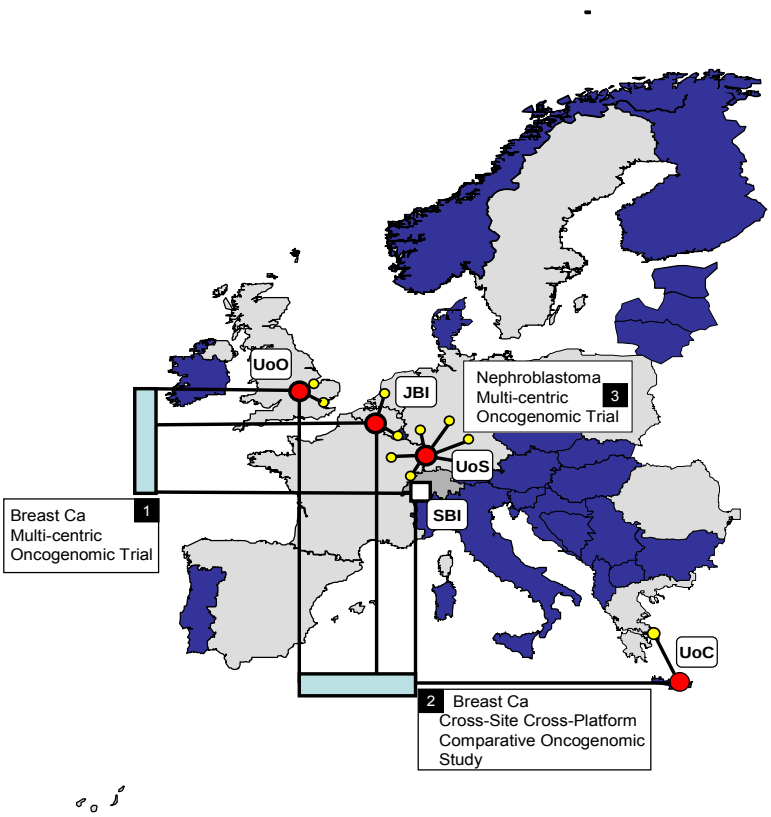


Figure 42: The ACGT clinical pilot sites

The ethics committee can give permission for the use of retrospective stored samples provided they are not linked to the patient directly. Also, we have been collecting tissues for patients signing their consent on the surgical consent forms. All projects have been submitted to the COREC (Research Ethics Committee) to obtain approval before proceeding with the research. As a condition of obtaining approval the security information, links to patients, how the information will be used and confidentiality need to be described and approved by the ethics committee.

Informed consent for tissue donation is obtained by giving the patient a written information sheet which they have time to consider for at least 24 hours before being requested to sign the consent form. This consent sheet is approved by the local ethics committee as part of the overall ethics application. The Oxford group in ACGT has recently had an MHRA audit relating to patients in clinical trials which was approved.

Samples are given an accession number which can link a pathology number to the sample, but when samples are handled all that is known is the accession number so the scientists and statisticians analysing the sample cannot relate this back to the patient. A specific tissue

bank manager organizes the tissue banks and the allocation of samples for research, once ethics committee approval has been obtained.

Clearly, to be of value in terms of new markers, prediction of sensitivity, response to treatment and outcome, there must be a link to clinical follow up. This is maintained in a separate database in a separate hospital and routine follow up on all patients with cancer is obtained for audit purposes in clinical management. These are maintained on a separate database and organized by a separate data manager funded by the National Health Service. The patients' identifiers on this database will include their pathology number. Thus, it will be possible to link the biological variables analyzed in the laboratory, e.g. gene array, using the coded numbers to the pathology.

However, this would not be done by the scientist doing the assays but by the tissue manager. It would be possible to link the data then to outcome, but this is from a separate database and would be linked by the clinical data manager in the National Health Service. No patient names would be used but it would then be possible to link the biological assay to the outcome. The ethics committee has approved this method of confidentiality and security. In the two sites where the databases are kept there is a backup using University or NHS computers and access is through password controlled computers in locked offices in swipe card controlled areas.

As new projects are undertaken then the new project needs to be approved by the ethics committee. There is a time limit on all projects and if a project goes beyond the time then an extension needs to be requested from the ethics committee or the project must halt.

This is bound by the national guidelines on research ethics committees published by the British government and which follows the new European regulations.

The standard form of consent for research trials and sample collection does involve requesting agreement individually for issues such as collaboration with other centres, lack of financial reward to the volunteer, access to relevant government agencies for inspection, collaborations in the future, right to withdraw consent at any time in the future, the knowledge of who is carrying out the research and how to contact them if they want advice on it, the agreement is that if relevant medical information became available form this it could be passed on to their practitioner. These are standard recommendations from the Medical Research Council and followed by all ethics committees in the UK in obtaining tissue consent.

Germany

In Germany, clinical trials are ruled by the German Drug Law (Arzneimittelgesetz). Of special importance are Articles 40, 41 and 42, which ensure the protection of human beings involved in clinical trials. The law has been amended last year, now allowing under very special conditions that children can be enrolled into clinical trials. This is also the case if the trial is not of direct benefit for the child involved, but possibly for the group the child belongs to.

The amendment also implements the Directive 2001/20/EC of the European Parliament and the Council of 4 April 2001 (good clinical practice in the conduct of clinical trials) into German law. According to this law, all clinical trials have to be approved by an official ethics committee. In order to gain approval, consent of the patient or – in the case of children – his or her legal representative has to be obtained after thorough information about the goals, procedures and possible side effects of the treatment to be tested.

Genetic data which are derived from patients are regarded as medical data, and therefore as sensitive data, which require protection equivalent to that of other sensitive data. In Germany, data protection in the private sector and concerning national public bodies is governed in general by the German Federal Data Protection Act (Bundesdatenschutzgesetz, BDSG), which entered in force on May 23, 2001. By this act, the old data protection law was amended, and with these amendments the provisions of the EU Data Protection Directive 95/46/EC of October 1995 have finally been implemented into national law. This law also contains many provisions, amongst others on the transfer of personal data abroad, as well as collection of data for research purposes.

Personal data such as names, birth dates and addresses collected for research must only be used and processed for this purpose. They should be anonymized as soon as the intended purpose allows it. In the meantime, personal data must be stored separately from medical information, which should not be linked to personal identifiers, but to a code instead.

When personal data are collected in the context of clinical trials, it is required according to German law which rules the conduct of clinical trials on medicinal products for human use (Arzneimittelgesetz) that patients give informed consent not only to the clinical trial, but – separately – also to the collection, storage, processing, transfer and analysis of personal data. Consent is legally effective only if it is given voluntarily and the subject has been informed of the purpose, nature, significance and implications of that use. The subject must know what he or she is agreeing to.

As a rule, information on the specific research project is necessary. However, restriction to a specific purpose may give rise to problems when blood or other tissue samples are collected prospectively for research purposes and stored in "biobanks". As infrastructure facilities for an indefinite number of research projects, they are unsuitable for their purpose if consent is to narrow. Therefore, a more broadly worded consent has been accepted by many ethics committees in Germany.

However, up to now, there is no specific regulation in place with respect to sample and data collection in biobanks. Up to now, several bodies have issued opinions on the use of human biological samples, amongst other the German National Ethics Council (March 17, 2004 "Biobanks for research", http://www.ethikrat.org/_english/publications/opinions.html) – by the way, in close cooperation with the French National Ethics Council (No 77 – March 20, 2003: "Ethical problems raised by the collected biological material and associated information data: 'Biobanks', 'Biolibraries'." http://www.ccne-ethique.fr/english/start.htm).

Currently it is debated in Germany – but certainly not only there – that it could be helpful to have an independent trustee who holds the key which provides the link between personal data and medical information. As a model case a data processing infrastructure has been established by the pharmaceutical company Schering, which includes different coding steps and an independent, third party trustee, providing a high level of privacy protection throughout the research process. Unfortunately, up to now there are only publications in German available which describe this model (http://www.tembit.de/fileadmin/PDFs/Datenschutz_in_der_pharmakogenetischen_Forschung_-_eine_Fallstudie.pdf). Similar considerations may become relevant for the ACGT-project in the future.

Regarding the transnational transfer of samples there are also no clear specific legal regulations available. But it is generally accepted, if the patients are informed about the fact that samples could be handed over to researchers in other countries, and they have consented to it. When the identity of cooperating partners is known at the time of data

collection, the patients should be informed about this. Whether patients could also consent to transfer to unknown partners, has to be examined.

According to the German National Ethics Council, such broadly framed consent must, however, be offset by a requirement that the samples and data, if they cannot be anonymized, may leave the area of control of the biobank only in coded form, except in circumstances provided by law. Personal data must not be passed on to third parties. In cases where external researchers require additional relevant data on subjects for their research, the data may be supplied only by an officer of the biobank to which the donors originally entrusted their samples and data, so that the external workers cannot identify individuals. Furthermore, full records should be kept of any transfer to third parties, to maximize transparency and to ensure that donors can withdraw their samples and data at any time. Donors' rights of withdrawal must be guaranteed whenever samples and data are transferred.

Belgium

The clinical partner from Belgium is the coordinator of the TransBIG project, aiming at translating molecular knowledge into early breast cancer management. TransBIG is partially funded by the European Commission under its Framework Programme VI.

The clinical research undertaken will fully obey existing national, European and International regulations. Patients participating in the research will be previously fully informed about the scope of the research and will be asked to give their explicit and written consent about it, as this required by currently legislation which is mentioned below. Tumour samples will be "leftover" tumour breast tissue obtained during diagnostic or therapeutic procedures. Blood samples will be, in the majority of cases, extra samples collected for research purposes only. Both tissues and blood samples will come from women suffering from breast cancer enrolled in a clinical trial run through the ACGT network. All human tissues are used for the identification of prognostic and predictive molecular markers. This implies that no hereditary genetic research is planned. What is to be examined is whether the human biological material may or may not be predictive of the efficacy of a specific treatment in each individual patient.

The Project will comply with the Law of August 22, 2002 relative to the patient's rights.

ACGT research in Belgium will fully comply with the Belgian Law regulating the operation of hospitals (7 August 1987 6), the Royal Decree (R.D.) n° 78 of November 10, 1967, the R.D. of August 12, 1994, the R.D. of 23 October 1964 that sets the standards to which the hospitals and their services must comply with.

The clinical trials are subject to the procedures of the Ethical Committees as these are regulated by the existing law on drugs of March 25, 1964 (that requires that a favourable opinion of an ethics committee is obligatory before the beginning of any clinical trial), as it was modified by the law of the 24 of December 2002 and complies with the Directive 2001/20/CE. It must be also mentioned that the recommendations of the Advisory Ethics Committee of Belgium (No 23 – 8/9/2003 – relating to the Ethics Committees) will be followed (the Law of May 7, 2004 rules the experiments on the human being, executed by the R.D. of June 20, 2004).

Clinical research will also comply with the directives of the "Conseil National de l'Ordre des Médecins" concerning research on human subjects, issued on the 22nd of August 1992, 17th of February 1996, 13th of December 1997, 19th of September 1998, 24th of April 1999, 15th of January 2000 and 19th of February 2000, and 19 2000. Of particular importance will also be

the recommendations of the Advisory Ethics Committee of Belgium (No 13 -19/7/2001 - relating to experiments involving human subjects), (No 2 – 7/7/1997 – concerning the convention of the human rights and biomedicine of the Council of Europe).

Clinical and genetic data collected are subject to the legislation about personal data and in particular the law of December 8, 1992 relating to the protection of the private life with regard to the processing of personal data , the Royal Decree (R.D.) of the 13.02.2001 (M.B. 13.03.2001) executing the Privacy Law. Furthermore related to the management of clinical information is the R.D. of December 15, 1987 and the R.D. of May 3, 1999 which determines the minimal general conditions for the medical file of an individual.

In case of problems with any new legislation relating to human biological sample collection and transfer, the Ethical-Legal Committee will evaluate the situation and will take appropriate action.

Greece

In Greece storage and processing of sensitive personal data is primarily governed by the following legislation:

- Law No 2068 /1992 validating the European Convention 108/1981 for the Protection of Individual from the Automated Processing of Personal data in Strasbourg 28th of January 1981.

- Greek Law No 2619 /1998 validating the Convention for the Protection of Human Rights and Dignity of the Human being with regard to the Application of Biology and Medicine: Convention of Human Rights and Biomedicine (Oviedo, 4th of April 1997) and in particular related to the issues raised by the EC Chapter II Consent, Article 5 (general rule), Article 6 Paragraph 2 (consent about children), Chapter IV (Human Genome), Article 12 (genetic examinations able to predict), and Chapter V (Scientific Research) Article 15 (general rule), Article 16 (Protection of Individual subjects to the Research) and Article 17 (Protection of Individuals Unable to Consent to the Research).

- Greek legislation about the consent to diagnostic practice is regulated by Law 2071/1992 and in particular Article 47 paragraphs 3, 4 and 5.

- According to the Greek legislation genetic data are “sensitive personal data” and in that sense protected according to Law 2472/1994 and Law No 2068 /1992, in compliance with Directive 95/46/EC on the protection of personal data.

- Laws about the Modernization of the National Health System article 57 for the Greek laws 2519/1997 and 2071/1992 article 57 regulate in particular the rights of patients, laws 2889/2001 (article 2 and 5) and 2071/1992 (article 61) the operation of the local ethical committees in hospitals and 2071/1992 (article 62) the code of medical practice.

- The Greek Drug Organization operates a National Committee for Clinical Trials according to ministerial decision 89292/2003 in compliance with Directive 2001/20

- Of specific importance are the Recommendations of the National Bioethics Committee, operating according to Law 2667/1998 for the collection and management of genetic data (2002) and the recent recommendation for the operation of review ethics committees for biomedical research (2005).

- Directive 2002/58/EC on privacy and electronic communications sector (Draft legislation to incorporate this directive into Greek Law is to be discussed soon before Parliament).

The processing of sensitive personal data is generally prohibited. By exception, the processing and recording of sensitive medical data is allowed, provided the Greek Data Protection Authority grants the required authorisation and one or more of the following requirements exists:

1. the data subject has consented in writing to the processing;

2. the processing is necessary for the preservation of the data subject's vital interest;

3. the processing concerns health related issues and is executed by a person who provides by profession medical services and is subject to a confidentiality duty or to related codes of conduct provided that the processing is necessary for the medical prevention, diagnosis, cure or management of health services;

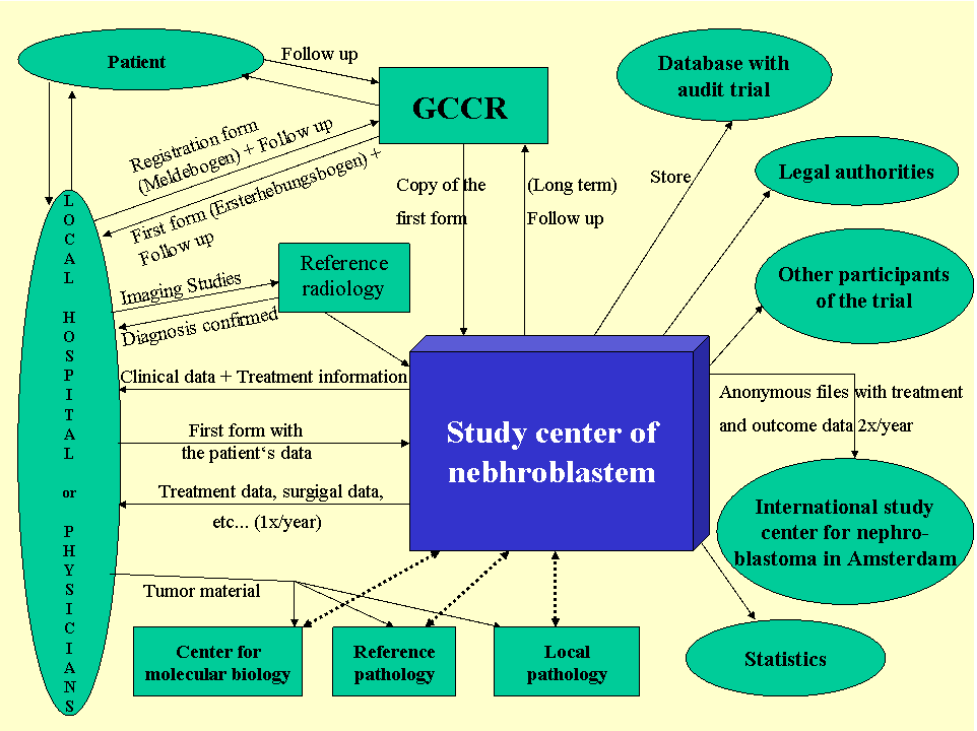
4. the processing is executed exclusively for research and scientific purposes and provided that anonymity is secured and all the necessary measures for the protection of the rights of the individuals to which the data refer are taken.

Of relevance is also the fact that the nationally funded research project "Prognochip", focusing on clinico-genomic breast cancer clinical study (which is a predecessor to the clinical pilot foreseen in ACGT) has been reviewed and approved on the 8th of October 2003 by Ethical Review Committee the International Agency for the Research on Cancer (IARC) applying in it assessments the International Guidelines for Ethics Review of Epidemiological Studies" (CIOMS1991) and the "International Ethical Guidelines of Biomedical Research Involving Human Subjects (CIOMS2002).

Especially with respect to the conduct of clinical trials, it would be useful to get more information.

21.3 Scenarios

To ensure the compliance of the clinical trials within ACGT with all relevant legal and ethical issues it is of high importance to identify, qualify (from a legal point of view) and structure the data flows that are produced during the patients' therapy. The clinical pilot trials are characterised by a multitude of data flows between different institutions.



The first step in the context of ethical and legal issues was to identify all data processing accruing in these clinical trials. The following figure gives an overview of the current practice, using the Nephroblastoma trial as example:

The figure shows the complexity of data flows. Based on the condition that each processing of personal data needs a basis of authorisation, it will be necessary to process as little personal data as possible. Hence the above shown data flows have to be divided into data flows that need to be personalised and into data flows that can be anonymous, as data protection legislation is not applicable to the processing of anonymous data.

The analysis of data flows showed that most of the data is processed pseudonymously, as the identification of each patient has to be guaranteed in order to give one the best therapy. Rendering data pseudonymous means replacing a person's name and other identifying characteristics with a label, in order to preclude identification of the data subject or to render such identification substantially difficult.

The processing of pseudonymous data needs as well as the processing of personal data a basis of authorisation. Still most of the involved parties only process pseudonymous data without having the link to the individual. Therefore we analyse the meaning of anonymous data in the European context with the primary goal to classify some pseudonymous data as anonymous data, so that a basis of authorisation is no longer needed.

A further step will be to identify one or many data controllers within the figure above. The data controller is responsible for the legitimate data processing. He can delegate the processing. Where other bodies are commissioned to collect, process or use personal data, the responsibility for compliance with the data protection provisions rests with the data controller. For the data transfer between the data controller and the data processor no basis of authorisation is needed. In other words the less data controller there are within ACGT, the easier it gets to process data within this project and the more difficult is it for the data controllers to ensure compliance with the legal framework of data protection.

21.4 Conclusions

To be able to ensure the full compliance of ACGT with all relevant legal and ethical issues, in-depth knowledge of the existing legislation, both at national, European and international levels has been acquired as a first step.

Based on this knowledge, a detailed analysis of all issues that may potentially arise in conjunction with the subjects and research activities of this project will be carried out, and necessary measures will be taken to conform to all the requirements.

The legal and ethical issues directly related to the research performed by ACGT that will be given attention throughout include the following:

- Patient Information
- Patient consent;
- Data protection;
- Transfer and storing of data and tumour/ blood samples;
- Scientific and Ethical review of the ACGT research projects.

Concluding, the ACGT Consortium wishes to confirm that the research does not involve:

- research activity aiming at human cloning for reproductive purposes,
- research activity intended to modify the genetic heritage of human beings which could make such changes heritable,
- research activities intended to create human embryos solely for the purpose of research or for the purpose of stem cell procurement, including by means of somatic cell nuclear transfer,
- research involving the use of human embryos or embryonic stem cells with the exception of banked or isolated human embryonic stem cells in culture.

On the other hand the research in ACGT involves biological samples, and other sensitive personal data, and also one of its pilots (nephroblastoma - Wilms' tumour) involves children.

The ACGT Consortium wishes to re-emphasize that:

- ✧ It is fully aware of European and National regulation related to its research activities and confirms that all such regulations will be observed. ACGT also plans, through its workplan, to assist the harmonisation of legislation related to trans-national clinical trials ethical and legal provisions;

- ✧ The Commission will always be kept informed that regional or national ethics approval has been obtained before the research to which it relates is carried out;

- ✧ Patient informed consent will be obtained and details of the information will be provided to patients;

- ✧ Specific, advanced methods for data storage and handling will be put in place (taking into account the Grid related dimension of the project) for ensuring patient data protection and confidentiality.

22 Security related issues

22.1 Introduction

Security is a concept that can be broadly defined to be “free from risk or danger”. In this first section this definition will be refined towards computer security and finally information security. The terms used in security related context are sometimes confusing and difficult to understand, therefore an attempt has been made to explain these in laymen’s terms and the more technical issues are skipped.

Computer security can be divided in 5 distinct domains:

- **Physical security**: In this domain the actions, movement and whereabouts of people and materials are controlled. The physical security domain also contains protection against the elements and natural disasters.
- **Operational/procedural security**: Everything from management policy decisions to reporting hierarchies is part of this domain.
- **Personnel security**: The hiring of new employees, doing background screening and monitoring is part of Personnel security. Also training of staff, security briefings and how to handle the departure of an employee.
- **System security**: Access controls to a computer system, authentication and assignment of privileges to users. Maintaining system and file integrity; backups, monitoring, logging and auditing of computer systems are all part of the domain.
- **Network security**: Protection of network and telecommunications equipment, servers and transmissions. Preventing and detecting intrusions, firewall policies, …

Security should be taken seriously in all these domains to get an optimal protection of resources and information. Most of these domains fall outside the scope of this document, which focuses on information systems.

22.2 Information systems security

For this document the term security is considered to be information systems security, which covers all aspects of data or information protection. The U.S. National Information Systems Security Glossary defines Information systems security (INFOSEC) as:

> *The protection of information systems against unauthorized access to or modification of information, whether in storage , processing or transit, and against the denial of service to authorized users or the provision of service to unauthorized users, including those measures necessary to detect, document, and counter such threats.*

There are 4 fundamental attributes of information security. The CIA triad, Confidentiality, Integrity and Availability, extended with Accountability.

- **Confidentiality**: “Restrictions on the accessibility and dissemination of information.” or “ensuring that information is accessible only to those authorized to have access”[18]

- **Integrity**: “The quality of correctness, completeness, wholeness, soundness and compliance with the intention of the creators of the data. It is achieved by preventing accidental or deliberate but unauthorized insertion, modification or destruction of data in a database.” Source integrity is having the assurance that the sender of the information is who he claims to be.

- **Availability**: “The accessibility of a system resource in a timely manner;”

- **Accountability**: “Clear accountability involves the processes, policies and controls necessary to trace the actions to the source. Accountability directly supports non-repudiation, deterrence, intrusion detection, recovery and legal admissibility of records”

A set of technical and non-technical solutions exists to protect these attributes of information. For each of these attributes a short list of topics is discussed.

22.2.1 Authorisation (Access Control Model)

The topic of **authorization** handles all the issues related to access control of resources in the most generic model. Well known examples of access control concepts are: Discretionary Access Control (DAC) where the owner of a resource can pass access rights directly on to another entity, Mandatory Access Control (MAC) where users do not have full control over the resources they own. In MAC (typical for the military) the user is limited in granting access to the security policy setup by the system administrator and cannot grant less restrictive access than that defined in the policy. In a Role Based Access Control (RBAC) system users can assume certain roles in the system. Access control is then decided not on individual identities but on the roles that can be assumed by the users. This allows for a much more fine grained and manageable access control system than for example DAC. A more advanced form of access control is Policy Based Access Control. Complex policies define the access decision and can for example take the context of the data access into account. Such policies are usually described in XACML, which will be described shortly in the section on Open Source tools.

.A resource[19] is protected by an enforcement component[20] that will grant or deny access to an actor[21] according to the decision made by a decision component[22] based on a policy[23] and identity information and extra information on the request and context[24] .

---

[18] ISO definition

[19] Both the ISO definitions as the XACML wording are used extensively. For the sake of readability both terms are given here but can be used interchangeably

ISO: target, XACML: resource

[20] ISO: Access Control Enforcement Function (AEF), XACML: Policy Enforcement Point (PEP)

[21] ISO: initiator, XACML: agent

[22] ISO: Access Control Decision Function (ADF), XACML: Policy Decision Point (PDP)

[23] ISO: Access Control Policy, XACML: Policy Document

[24] ISO: Access Control Decision Information (ADI), XACML: (certificate) attributes

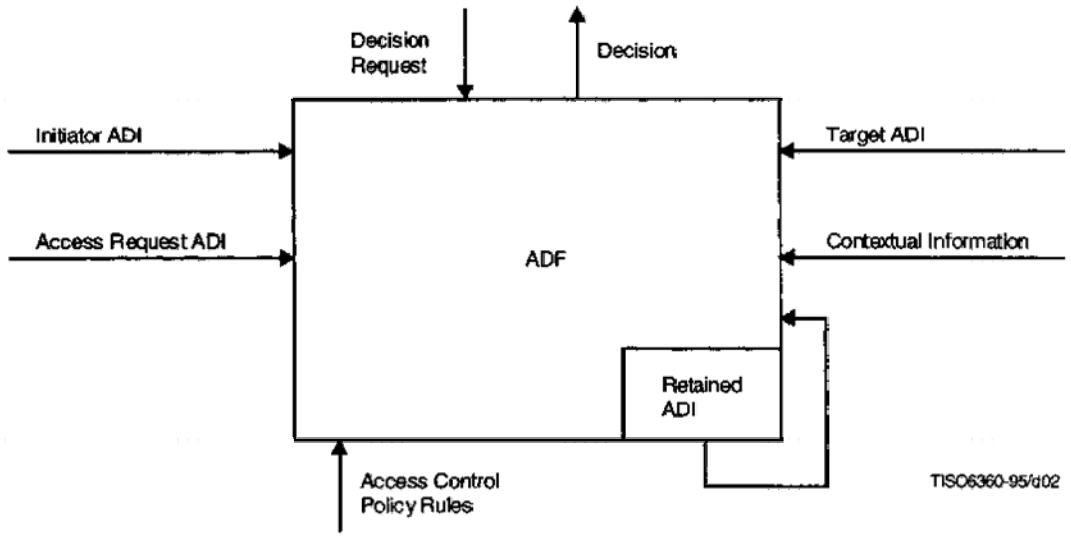
The **target** itself can take any form, ranging from a file on a computer, over a record in a database or the full database itself to computer processing time in a server cluster. Even the computer hardware itself can be a resource.

The **enforcement component** is an actor or process that accepts identity information in a well defined form from the agent and based on the policy document grants or denies access to the resource.

The **actor** is a user or process that wishes to access the resource. This can be a real person or computer program acting on behalf of that person. To allow access control this actor needs to be identified in manner acceptable to the enforcement component.

A **decision component** makes the actual decision whether or not to grant access to the resource at the time of the request. In a policy based access control system the range of information used by this component can be rather extensive. The identity information provided by the actor, information on the resource and one or more policies describing access rights to the resource, but also context information such as local time of the server or previous access requests for that resource.

The **policy** describes the conditions under which a user can access a resource and what the user is allowed to do with the resource. These conditions often contain restrictions regarding identity information but can also restrict access based on the source of the identity information, date or time, the action the agent wants to perform on the resource. For a single resource more than one policy can be defined in which case all documents are valid and the strictest result is taken. All policies need to grant access for the given conditions.



The **identity information** an agent provides to the authorization mechanism needs to be in a form that is accepted by the policy enforcement point. An **authentication** mechanism is a process that verifies the claims of identity by the actor. The actor can claim an identity in different ways. Knowledge of a password is for some services a strong enough claim of identity, while other services require a stronger claim such as time limited, digitally signed certificates. After authentication it is often needed to transform the identity information to a format that is accepted by the decision and enforcement component.

In today's practice, the components involved in access control are usually not clearly separated and can be situated on one single machine or even program. In more complex

situations each component can be a service located on a distinct server. Whatever form of implementation is used, the components need to be able to establish a certain level of trust with each other.

Trust itself is a difficult subject in computer security, a viable definition is: “an entity can be trusted if it always behaves in the expected manner for the intended purpose”. The trust issue covers all relationships within a (distributed) computer system. The goal of security systems is to provide a technical implementation of assuring trust (relying on cryptography).

22.2.2 Integrity

Data Integrity can be compromised either deliberately or accidentally. An accidental breach of integrity can be caused by the loss of network connectivity, a database crash during a transaction or simply the accidental removal of a file by a system administrator. Accidental integrity compromise affects normal operations and the mechanisms to prevent or recover from accidental integrity compromise; such as backups, data retransmissions and control bits will not be discussed further in this text as they are not specific to security.

A deliberate breach of data integrity could be an attempt to change an instruction sent to a person or computer, changing a database in order to get access or to deny access to other users or to change data in a way favourable to the attacker. Or in a case where the attacker simply wants to cause damage, corrupt the data in a random manner.

Source integrity considers the claim that the source of the data is who it claims to be. On an insecure medium, there is no guarantee that received data is authentic. An attacker can place fake messages on the medium and claim to be from a trusted source. Message integrity considers malicious alteration of content. Protection of integrity heavily relies on cryptographic methods such as Digital Signatures or Message Authentication Codes. Note that it is usually not possible to prevent any tampering of the data, but at least it can be detected.

22.2.3 Trust

Trust is one of the most difficult problems in security, and spans a number of different issues. The primary issue is “How can two parties establish a trust relationship, without ever having met?”. The solution commonly used assumes that trust is transitive (which can be disputed). If entity A trusts entity B and entity B states that he trusts the entity C then entity A can put a certain level of trust in the entity C.

Technical implementations are based on public key cryptography (implemented in Public Key Infrastructures (PKI)). Trust relations are expressed in certificates or assertions, which are nothing more than “sworn statements about a party” signed by authorities. These Authorities usually form a hierarchical structure, like in the standard known X.509 hierarchical PKIs. Other approaches such as “web of trust” do not distinguish specific authorities as the base of trust, each entity is considered equal.

22.2.4 Availability

Guaranteeing availability for a service is a very difficult task. As with integrity the availability can be lost either due an accidental or deliberate action. Accidental actions or situations, such as human errors or hardware failures are not specific to computer security.

The chance that a service becomes unavailable can be limited by redundancy. When one service is on standby and takes over the task as soon as the main service becomes unavailable, then a non deliberate cause of failure (human error, hardware failure, etc) will have very limited effect on the overall availability.

An attacker has different options to make a service unavailable, and does not have to breach the security of the targeted system to make it unavailable. The term Denial Of Service (DOS) attack denotes all types of attacks (without intrusion) that cause a loss of service to users, typically the loss of network connectivity and services by consuming the bandwidth of the victim network or overloading the computational resources of the victim system.

A denial of service attack is the most difficult attack to counter. Although a number of security measures can be taken to make it hard for an attacker to launch a DOS attack, in the end, it is a matter of resources. With a Distributed Denial Of Service attack (DDOS) it becomes impossible to single out the source of the attacker to cut it off because the attack comes from many seemingly random computers all over the Internet. Computers of home users and businesses all over the world have been taken over by hackers without the knowledge of the owners and entire networks of hacked computers are available to start such an attack.

It is interesting to note that security measures can make a service more vulnerable to DOS attacks. As a simple illustration: a system with account locking, where an account gets disabled after 3 successive authentication failures can easily be shut down by an attacker.

22.2.5 Accountability

The most important part of accountability is the ability to create an Audit Trail, to observe and chronologically log all actions undertaken by users. The created audit trail needs to be secure, the integrity of logs should be maintained at all times. If there is a suspicion that a log entry can be altered, deleted or that extra entries can be added to the audit trail at a later time then the whole audit trail cannot be trusted and an actor can deny actions recorded in the logs. Accountability relies on accurate, extensive and secure logging and thus relates to the trust problem. Accurate timekeeping at all points of the network is thus an absolute must.

Strongly related to accountability is non-repudiation (in some sense it could be considered a synonym), which has been defined as follows:

> *Non-repudiation is a property achieved through cryptographic methods which prevents an individual or entity from denying having performed a particular action related to data (such as mechanisms for non-rejection or authority (origin); for proof of obligation, intent, or commitment; or for proof of ownership)*

Non-repudiation or accountability is a necessary component of information security. While it does not prevent breach of information security in a direct sense, it does allow quickly tracing back and identifying the source of a malicious act. By identifying the source's future malicious acts by that source can be prevented and the source can be held liable for its actions.

22.3 Privacy Enhancing Techniques

22.3.1 Introduction

Information security is just as any form of security a matter of raising the barrier for a possible attacker. A locked door will pose a larger problem for a burglar than an unlocked door, but a determined burglar will have no problem entering a building. An alarm system raises the barrier yet again but is not infallible. A guard dog, cameras, motion detection, all these help raise the barrier for a potential burglar, yet they all have a weakness that can be exploited.

In computer security a similar set of measures can be taken to prevent intrusion and theft of data, yet no security measure is infallible. Whether the breach of security is accidental or deliberate, the data may be lost or spread in the public. For medical information it is extremely important to prevent that data is leaked into the public.

A technology which is designed to safeguard privacy is generally referred to as Privacy Enhancing Techniques (PETs). The concept of PETs has been around since at least the mid-1990s. Marc Rotenberg has tracked down an early usage of the phrase (at that stage without the acronym) in a CPSR (Computer Professionals for Social Responsibility) Statement of 10 June 1991[25].

Privacy Enhancing Technologies (PETs) can be defined as:

> *"A coherent system of ICT measures that protects privacy by eliminating or reducing personal data or by preventing unnecessary and/or undesired processing of personal data, all without losing the functionality of the information system."*

The broad world of privacy protecting techniques can be split up into two main classes. The first type of PETs can provably eliminate or minimize the collection of personally identifiable data while still enabling someone to engage in a transaction or in communication (sending mail, doing a purchase, surfing the web, etc.). Basically these techniques protect a person's privacy when that individual is engaging in an operation. In such scenarios, the person whose privacy is to be protected, remains more or less in control of the situation. Examples of these PET are anonymous remailers, digital certificates with limited information, digital cash, pre-paid smartcards (e.g. cell phones)

Another class of Privacy Enhancing Technologies and Privacy Supporting Techniques is usually not under the control of the person to be protected. They are used for protecting identities in data collection. The goal is to safeguard the privacy of subjects of data gathered, without limiting the useful information content of that data. It is that set of Privacy Enhancement Techniques that is relevant for ACGT.

[25] The term "Privacy Enhancement" however, was used at least as early as the mid-1980s, in the RFC series 989 (February 1987), 1040 (January 1988), and 1113-1115 (August 1989), which defined a 'Privacy Enhancement for Internet Electronic Mail". The term referred, however, only to the narrow concept of message transmission security, and its requirements of confidentiality, authentication, and message integrity assurance

22.3.2 Privacy Enhancement Techniques

The following paragraphs give a short introduction on several types of PETs which might be useful at some point in ACGT. The given list of PETs is not exhaustive and is meant as an introduction. Discussing all methods would lead too far and is not in the scope of this document. WP10 together with WP11 will further elaborate on this.

Encrypted storage: A straightforward method of ensuring privacy is by encrypting the storage of the data, an entire file system, a single file or a single record in a database. Encrypted storage prevents access from unauthorized persons by requiring the encryption key to access the data. This method has some as of yet (unresolved) issues. Key management for the encryption is a difficult problem: should there be a different encryption key for each file, for each record in the database? How does one perform a search over encrypted storage? Indexing is difficult and could leak information on the database. Operations on the database are slowed down because all the data needs to be decrypted when read and encrypted when stored. And how does one handle access revocation? Access is granted by providing the encryption key to the user, once given it cannot be taken back. Do all encrypted files need to be re-encrypted with a new key each time access is revoked? No optimal solution exists, different approaches make different tradeoffs.

There are many tools and hardware devices available for encrypted storage, ranging from home users to large companies. The popular GPG [WEB1] is a free tool available on most systems that can be used for encrypting and signing data. For databases integrated encryption tools [WEB2] can be used or the data can be encrypted before storage in the database, using for example dedicated applications [WEB3]. Encrypted file system solutions [BIN1999, FU1999, MIL2002, GOH2003, KAL2003] exist in many forms and abilities, some modern operating systems support encrypted file systems by default [WEB4] while others rely on 3rd party programs [WEB5].

Custodix commercially deploys and further develops a “Privacy Enhanced Storage Framework” allowing fine-grained sharing of encrypted data. Current implementations are directed towards Electronic Data Capture web applications. The core component (Key Management Service) could be relatively easily ported to the Grid, offering shared encrypted storage in the Grid.

De-identification of data: De-identification data is generally coined anonymization. Often the term merely refers to stripping off the identifiers of a dataset, or “coding” them in such a way that the nominative identities can no longer be recovered. The latter allows the individual to be tracked in time or space.

It is clear that this alone does not solve the privacy problem. Truly breaking the link between the identity of an individual and data about that person is far from straightforward. The “direct link” is easily broken: direct nominative identifiers such as name, address, telephone number, can easily be removed or transformed. The most interesting data for mining however often contains information which can be “indirectly linked” to a single individual. This means that through inference, the identity of data subject can be discovered. Herein lays the difficulty in truly making data records anonymous.

- **Direct re-identification (one could call these “first order” techniques)** By examining a data record one can immediately and with certainty retrieve the identity of the person.
  e.g. the de-identification scheme can be broken (reversed); the privacy protection system failed to remove several nominative identifiers (e.g. address and age)

- **Indirect re-identification (one could call these "second order" techniques)** Based on information hidden in a data record (probabilistic narrowing down), or based on Correlation between different data records.
  e.g. The search space to which the person belongs can be narrowed down by exploiting certain known characteristics, a little amount of extra information can then identify that person; From genomic data, a unknown person could be known to have brown eyes, blond hair, blood group A+ and a particular uncommon genetic defect. If it is known that the specific record belongs to a person from a certain town, that person could be identified; Lineage included within a database; genetic fingerprinting and inheritance can lead to re-identification through association of several data records; Progress of a treatment (temporal study) can create a unique pattern on an individual.

The term pseudonymization refers to the replacement of identifiers by a pseudonym, a pseudo-ID. When the same pseudonym is used over different datasets then they can be combined in a meaningful way for analysis without any need of knowing the real identity of the subject.

A difficult issue in de-identification (actually in the general processing) of data lies in the fact that much of the available data is unstructured. Detecting identifiers in free text, images or audio streams is quite complex [WEB6, TAI2002]. Most de-identification implementations treat structured data only. One relies on the fact that the conversion from unstructured data is needed anyway for further proper data analysis [REC2003].

Statistical Disclosure & databases: Statistical Disclosure Control makes it possible to disseminate data from a private database without a too great risk to the personal information contained within. Instead of allowing queries on the data records only statistical results are returned. A statistical database can be defined as follows: "*An SDB has been defined as one which returns statistical information, such as frequency counts of records satisfying some given criteria, as opposed to a database which returns details of an entity, for example, name and address of an employee*" There are two mayor groups of disclosure techniques used in statistical databases.

- **Restrictive methods**: These methods all limit the amount of information that can be published. Extreme values can be grouped in top or bottom categories, several categories are collapsed into one so that no records stand out from the rest. Some values are simply suppressed for publication.

- **Perturbation Methods**: Values of some variables are modified in such a way that the statistical properties of the data does not change (too much). Values are rounded in a random fashion, data is swapped amongst records without changing the statistical properties of the whole dataset, and data records are combined and aggregated into groups. Insertion of fake data, or returning a completely different database with the same statistical properties as the original is also a method used for some applications.

The main goal of these data protection techniques is to prevent identification of "rare" or unique datasets with respect to a certain combination of characteristics.

Dynamic database protection: This contains a set of methods to protect privacy of subject data during database operations. These methods go beyond the standard access control available in DBMS (column based AC).

- **Query modifying:** Queries sent to the database are parsed and modified before they are passed on to the database engine. The modifications are done based on the access policy concerning the user executing the query. For example, to each query from a user who has access only to data that is older than 3 years, an extra clause can be added to only select old data. Changing the query beforehand is much more efficient than filtering the queried data afterwards because this method takes advantage of the optimizations of the database engine.

- **Privacy metadata (Hippocratic databases)** [AGR2002]: Some research has been done towards databases that take into account the intention of queries and data and decide on database level if the query is consistent with the policy and its intention. Data, tables, and queries are tagged with metadata which is then used by the engine when making the decision whether or not to return the result.

- **Query Set Restriction:** A dynamic form of (statistical) disclosure control. In this approach a query on a database filled with microdata is either allowed or denied based on some measure, in order to avoid leaking of information on individual records.

    - *Query set size control* works by setting lower and upper bounds [FEL1972] for the size of the query answer set based on the properties of the database and on the preferences fixed by the database administrator. If the number of returned records did not lie within these bounds, the information request would have to be rejected and the query answer is denied.

    - *Auditing* involves keeping up-to-date logs of all queries made by each user and constantly checking for possible disclosures whenever a new query is issued. One major drawback of this method is that it requires huge amounts of storage and CPU time to keep these logs updated.

Computation related PETs:

- **Secure Multi Party Computation (MPC)** [GOL1998]**:** This family of cryptographic protocols is designed to share information from private data and calculate a certain function on the data without disclosing the private data to all participants. A MPC protocol is dubbed secure if no participant can learn more from the description of the public function and the result of the global calculation than what he/she can learn from his/her own entry - under particular conditions depending on the model used.
  A typical application example is comparison amongst peers. A number of medical centres could for example want to compare their treatment performance for a specific disease without having to disclose their own success rates (in order to avoid public shaming or unfair advertising).

- **Privacy Preserving Computation:** Techniques that allow third party computation of data while not disclosing that data. These techniques often rely on specific encryption schemes.

Privacy Preserving Data Mining: This subset of data mining techniques is developed with privacy concerns in mind. Many techniques used in this area are similar to statistical disclosure methods and databases while the more cryptographically oriented methods lean towards secure computing. An overview of publications can be found at [WEB7], a state-of-the-art review including a classification in [VER2004].

22.4 Available open source tools and services

This section describes some existing Open Source implementations of security frameworks related to Grids, security tools and technologies and standards with Open Source implementations. This list concentrates on those deemed useful for ACGT and is certainly not exhaustive.

22.4.1 GLOBUS Grid Security Infrastructure

The open source Globus Toolkit [WEB8] is a fundamental enabling technology for the "Grid," letting people share computing power, databases, and other tools securely online across corporate, institutional, and geographic boundaries without sacrificing local autonomy. The toolkit includes software services and libraries for resource monitoring, discovery, and management, plus security and file management.

An important part of the Globus Toolkit is the Grid Security Infrastructure (GSI) [BUT2000, FOS1998]. The primary motivations behind the GSI are:

- The need for secure communication (authenticated and perhaps confidential) between elements of a computational Grid.
- The need to support security across organizational boundaries, thus prohibiting a centrally-managed security system.
- The need to support "single sign-on" for users of the Grid, including delegation of credentials for computations that involve multiple resources and/or sites.

Several modules have been developed for the Globus Toolkit to meet these requirements. Components for authorization, delegation and secure messaging, credential management and supportive tools are available and can be installed separately from each other.

Recently the focus is on Web Services and the technologies are being adapted to take advantage of standards for Web Services. A distinction is being made between pre-WS technologies and WS technologies. Following the WS-* standards Globus Toolkit [WEB9] has implemented mechanisms for

- Message level security
- Transport level security
- Authorization framework

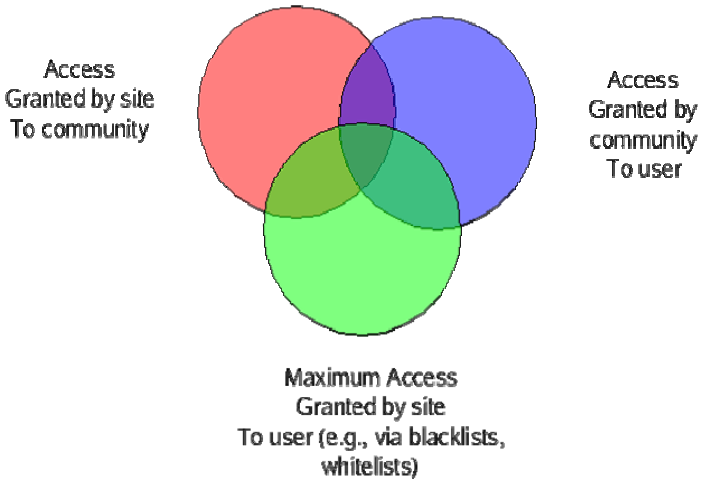
The pre-WS-* tools are described shortly after which the new authorization framework for GT 4 is discussed. As Web Services become more popular this document will put more focus on Web Service enabled technologies.

The tools described as part of the Globus Toolkit are part of the default installation package and will be installed together with the rest of the package. The latest version can be found online at “http://www.globus.org/toolkit/downloads/4.0.2/”. The individual packages can be installed separately from source, which can be found at the same location.

22.4.1.1 Community Authorization System (CAS)

The Community Authorization Service (CAS) is part of the Globus Security Infrastructure and is developed for non Web Services in a Grid. A CAS server issues credentials that are a combination of a proxy certificate [TUE2002] issued by the user and a signed policy

assertion of the CAS server. These credentials are used by the virtual organization users for obtaining access rights to resources (fine-grained). Servers recognize and enforce the assertions. CAS is designed to be extensible to multiple services.



The CAS system supports multiple policies on different levels: a policy describing the rights and obligations of the virtual organization regarding the resource, a policy of the Vitrual Organisation (VO) towards its members and finally a local policy at the level of the resource towards individual members.

The rights of an individual user are then the intersection of the rights given by the policies at each level.

To access a resource a user first connects to the CAS server and obtains a CAS credential, cryptographically signed by the server. The CAS credential is a combination of a (proxy) certificate issued by the user with the signed policy assertion issued by the CAS server. The user then presents the credential to the resource (a CAS enabled service) which then checks the validity of the credential and enforces the policies.

22.4.1.2 Delegation service

The Delegation Service is a component in Globus Toolkit 4.0. This component provides an interface for delegation of credentials to a hosting environment. This enables a single delegated credential to be shared across multiple invocations of services on that hosting environment. It also provides a means for credential renewal.

22.4.1.3 Credential Management

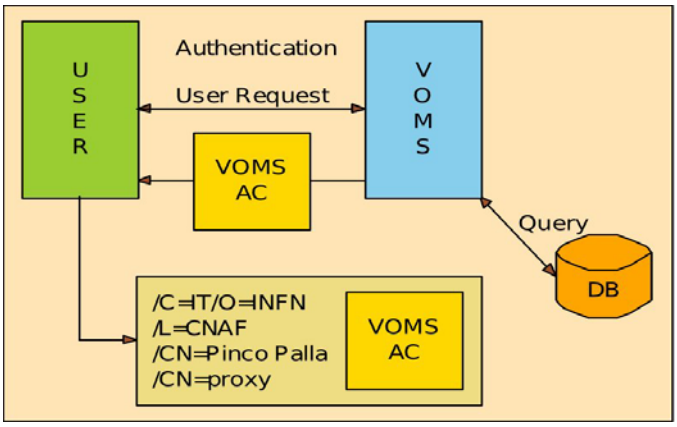
There are two main technologies used for credential management in the Globus Toolkit: Credential creation and a credential repository.

For credential creation a Certificate Authority is used. A very simple Certficate Authority called Simple CA has been created for test purpose only. Some more advanced CA tools are discussed in a following section.

The credential repository called MyProxy [NOV2001] is a repository in which one can store X.509 proxy credentials, protected by a passphrase, for later retrieval over the network. This eliminates the need for manually copying private key and certificate files between machines. MyProxy can also be used for authentication to Grid portals and credential renewal with job managers.

22.4.2 VOMS

VOMS [ALF2003] or Virtual Organization Membership Service performs authorization according to group membership privileges. A VOMS credential contains a non-critical certificate extension that is presented to the resource. In this way the relevant authorization information is packaged directly with the identity information used for authentication.



Users of a VO are organised in groups which, in general, form a hierarchical structure with the VO itself as root. Groups can have multiple ancestors but cycles are not allowed in the structure. A user being member of a group is automatically member of all its ancestor groups. Users have roles and capabilities in groups which are inherited from top to bottom. (If a user has role A in group X then he will also have role A in all subgroups of X where he is also a member, but not the other way)

The VOMS system attempts to let users present the authorization data as they try to access the resource, shifting form a pull model to a push model.

Normal operation of a VOMS server is as follows:

VOMS uses user's proxy certificates that contain authorization information from the VOMS server(s).

1. authentication between VOMS server and user
2. user sends signed request to VOMS server
3. VOMS server checks request on correctness
4. VOMS server sends signed reply as Attribute Certificate
5. user checks validity of reply
6. repeat 1-5 for all VOMS servers needed
7. user creates proxy certificate with all info from VOMS servers into non critical extensions
8. user can add its own authentication info (e.g. kerberos).

A VOMS server can be installed from source, which can be downloaded freely from the CVS repository. More information on accessing the source can be found at “http://infnforge.cnaf.infn.it/scm/?group_id=7”. A VOMS administrative interface is available either as binary installation package or from a CVS module.

22.4.3 Shibboleth

The Shibboleth [WEB10] software implements the OASIS SAML 1.1 specification, providing a federated Single Sign On and attribute exchange framework. Shibboleth also provides extended privacy functionality allowing the browser user and their home site to control the Attribute information being released to each Service Provider. Using Shibboleth-enabled access simplifies management of identity and access permissions for both Identity and Service Providers.

While Shibboleth was made for websites only, a project is in an advanced stage to integrate shibboleth with the Globus toolkit. The Integration requires that a plugin [WEB11] is installed for both the Shibboleth software as the Globus Toolkit.

When a user tries to access a resource he will come into contact with several parts of Shibboleth. At first contact the resource has no information about the user which is needed for authorization. The Shibboleth Indexical Reference Establisher (SHIRE) is a service that will try to get a handle for the user. To do this the user will be redirected to the Where Are You From (WAYF) service that allows the user to select his home site where authentication will be done. The WAYF is configured so that it knows the location of the Handle Service for each home site.

The Handle Service (HS) authenticates the user locally and provides a handle that can be used to request attributes. The handle in form of a SAML authentication statement is sent back to the SHIRE, together with additional information; e.g. the attribute provider where attributes on the handle can be requested.

The SHIRE then passes on the handle to the Shibboleth Attribute Requestor (SHAR) which sends a (SAML) attribute request message to the Attribute Authority listed in the handle. The response is validated and the embedded attributes are forwarded to the resource by the SHAR. The resource now has the information to make an authorization decision.

The strength of Shibboleth lies in the Single Sign On that it offers across domains and the protection of privacy since the local Attribute Provider only returns the attributes needed for the authorization decision for the specific resource.

The Shibboleth software can be downloaded from their main site at “http://shibboleth.internet2.edu/latest.html”. Note that the Service Provider software also has a C++ version next to the default java implementation.

22.4.4 PERMIS

PERMIS is an implementation of a Role Based Access Control infrastructure [WEB12] using X.509 attribute certificates to store users roles. A typical use of the PERMIS system is as follows:

System administrators will write one or more policies, specifying which roles have which privileges, and what kind of credentials will be recognized by PERMIS. The policy will be used by PERMIS for all reasoning regarding authorization. Attributed administrators will issue credentials containing users' attributes, telling what roles the users have. A request to access a resource is intercepted by PERMIS, the user's credentials will be analyzed, and only those that can be validated by the credential validation rules in the policy will be accepted. Then the PERMIS system will use the association of roles and privilege as specified in the policy, and the association of users and roles as specified in the recognized credentials, and the requested actions and the resource, to render an authorization decision.

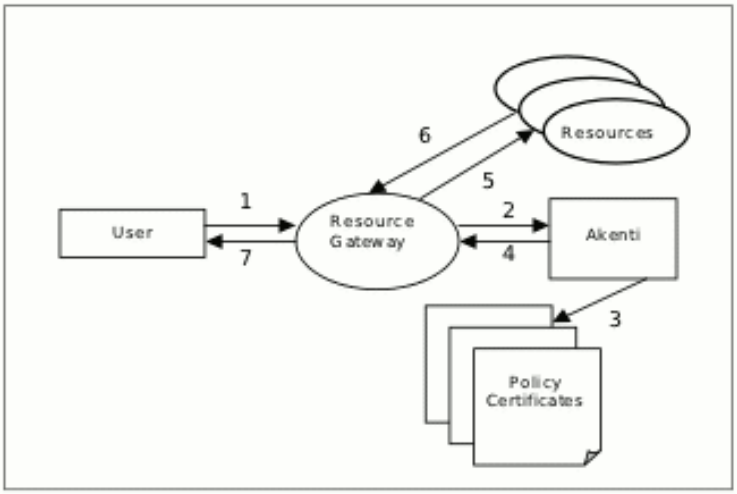
PERMIS supports two different flavours of SAML, firstly the openSAML implementation and secondly the Grid Authentication profile of SAML [WEL2004] as implemented in the Globus Toolkit version 3.3.

To install PERMIS as a service refer to the online download and documentation page at http://sec.isi.salford.ac.uk/permis/downloads/download.shtml. Binary releases and a short installation manual are provided on the site. Instructions on how to combine PERMIS with the Globus Toolkit or Shibboleth software are also provided.

22.4.5 Akenti

Akenti [THO2003, WEB13] is an authorization system that gives users a single Virtual Organization wide identity in the form of an X509 certificate. It is designed to allow multiple independent stakeholders to set the access policy remote from the resource itself. Akenti is a Policy decision Point that accepts X509 certificates as identification information and signed authorization policy certificates from multiple sites to make the authorization decision. The PDP service returns a signed capability certificate that can be used directly by a Policy enforcement Point (PEP).

Globus proxy [TUE2002] certificates are supported by Akenti and an apache server module (http://dsd.lbl.gov/Akenti/docs/mod_akenti.html) has been implemented.



At the moment of resource access a resource gateway contacts an Akenti server and asks what access the user has to the resource. The Akenti server fetches all relevant policies, verifies the origin of the policies, evaluates them and returns the allowed access. While CAS and VOMS work according to a push model, Akenti uses a pull model.

Akenti has been implemented both as Java as C software and can be installed through binary distribution or from source code. The software is distributed under a BSD license and can be downloaded from the Akenti site at http://dsd.lbl.gov/Akenti/download.html .

An integration module for the Globus Toolkit GRAM module can be found at http://dsd.lbl.gov/akenti/codeDist/GRAMAkentiAuthz.html, although the stability of this package is not guaranteed.

22.4.6 Anonymization tools

A number of tools and plugins that will anonymize existing data files. For the DICOM [WEB14] software, a set of tools for the medical imaging community an anonymization tool can be found at [WEB15] and some viewers with built in anonymization at [WEB16, WEB17]. For the anonymization of log files a few tools can be downloaded from [WEB18, WEB19].

True de-identification is however a complex process. Privacy risk analysis on datasets (a-priori and a-posteriori) is the most important, but also most complex part of the de-identification process. Tools such as u-Argus and Datafly are available for disclosure control of microdata. Further elaboration is out of the scope of this deliverable.

Custodix performs risk assessment and deploys anonymization tools as part of a Trust Service (many anonymization schemes require a Trusted Third Party).

22.4.7 Other General technologies and standards

The three most important technologies are discussed in more detail, SAML, XACML and WS-*. Besides these three technologies, a whole set of technologies and frameworks exist such as the Ponder Language, Cassandra, Keynote, Delegent, Kerberos, PRIMA, Cardea. These technologies are not always Open Source and others are outside the scope of this document. The three discussed are so important that they could not be left out.

22.4.7.1 SAML

The Security Assertion Markup Language (SAML) [WEB20] is an XML based framework developed by the Organization for the Advancement of Structured Information Standards (OASIS). The goal of SAML is to support interoperability between different services which require security services. SAML provides the technology for Single Sign On (SSO) between different security domains where users are authenticated in one domain and where their credentials are transferred to other partner domains without the need for the users to re-authenticate.

A SAML document is composed out of one or more assertions made by an authority stating facts about a subject. SAML 1.1 supports 3 types of security assertions: ***Authentication statements, Attribute statements and Authorization decision statements***. Other statements can be created using SAML extensions.

An Authentication statement asserts that the principal did indeed authenticate with an authentication service at a particular time using a particular method of authentication. Other information about the principal may be disclosed in an authentication statement, e.g. email address, age, or any other attribute that can be used by the authorization service to make an authorization decision.

An Attribute statement states that a subject is associated with a set of attributes and their respective values. Membership of a group or function within an organization is an attribute that can be asserted with an attribute statement.

An Authorization statement states which actions a subject is authorized to perform on a certain resource.

The example SAML statement asserts that user with email address user@mail.idp.org has authenticated using a password on 19/06/2002. The assertion is issued by the service at https://idp.org/saml/ and is valid for only 10 minutes.

```
<saml:Assertion
 xmlns:saml="urn:oasis:names:tc:SAML:1.0:assertion"
 MajorVersion="1" MinorVersion="1"
 AssertionID="..."
 Issuer="https://idp.org/saml/"
 IssueInstant="2002-06-19T17:05:37.795Z">
 <saml:Conditions
   NotBefore="2002-06-19T17:00:37.795Z"
   NotOnOrAfter="2002-06-19T17:10:37.795Z"/>
 <saml:AuthenticationStatement
   AuthenticationMethod="urn:oasis:names:tc:SAML:1.0:am:password"
   AuthenticationInstant="2002-06-19T17:05:17.706Z">
   <saml:Subject>
     <saml:NameIdentifier
       Format="urn:oasis:names:tc:SAML:1.1:nameid-format:emailAddress">
       user@mail.idp.org
```

```
      </saml:NameIdentifier>
      <saml:SubjectConfirmation>
        <saml:ConfirmationMethod>
          urn:oasis:names:tc:SAML:1.0:cm:artifact
        </saml:ConfirmationMethod>
      </saml:SubjectConfirmation>
    </saml:Subject>
  </saml:AuthenticationStatement>
</saml:Assertion>
```

The SAML protocol is a simple request-response protocol. SAML 1.1 defines just one *binding*, the SAML SOAP binding. A compatible SAML 1.1 implementation **must** implement SAML over SOAP over HTTP (a synchronous protocol).

Current implementations for SAML are:

- OpenSAML [WEB21]: a set of open source Java and C++ libraries that partially implement the SAML 1.0 and 1.1 specifications. Access to the source code is obtained through anonymous CVS access or through a web front-end for the CVS server at http://anoncvs.internet2.edu/cgi-bin/viewvc.cgi/opensaml/.

22.4.7.2 XACML

Another XML specification by OASIS is the eXtensible Access Control Markup Language [WEB22]. The policy language is used to describe general access control requirements, and has standard extension points for defining new functions, data types, combining logic, etc. In addition to a policy language, XACML specifies a request and response protocol for describing queries about a particular request and decisions made regarding the request. The response always includes an answer about whether the request should be allowed using one of four values: Permit, Deny, Indeterminate (an error occurred or some required value was missing, so a decision cannot be made) or Not Applicable (the request can't be answered by this service).

In a typical situation a subject (an agent acting on behalf of a person or a real person) wants to access a protected resource. Access is restricted by a Policy Enforcement Point (PEP) which executes the authorization decision of a Policy Decision Point (PDP). When a request for access reaches the PEP it will form a XACML request based on the authentication information provided by the subject and sends it to the PDP. The PDP will fetch the appropriate policies written in the XACML policy language and based on the policies determine the authorization decision according to the XACML rules for evaluating policy rules. The PEP and PDP might both be contained within a single application, or might be distributed across several servers. In addition to providing request/response and policy languages, XACML also provides the other pieces of this relationship, namely finding a policy that applies to a given request and evaluating the request against that policy to come up with a yes or no answer.

At the root of all XACML policies is a Policy or a PolicySet. A PolicySet is a container that can hold other Policies or PolicySets, as well as references to policies found in remote locations. A Policy represents a single access control policy, expressed through a set of Rules. Each XACML policy document contains exactly one Policy or PolicySet root XML tag. Because a Policy or PolicySet may contain multiple policies or Rules, each of which may evaluate to different access control decisions, XACML needs some way of reconciling the decisions each makes. This is done through a collection of Combining Algorithms. Each algorithm represents a different way of combining multiple decisions into a single decision.

There are Policy Combining Algorithms (used by PolicySet) and Rule Combining Algorithms (used by Policy). An example of these is the Deny Overrides Algorithm, which says that no matter what, if any evaluation returns Deny, or no evaluation permits, then the final result is also Deny. These Combining Algorithms are used to build up increasingly complex policies, and while there are seven standard algorithms, you can build your own to suit your needs.

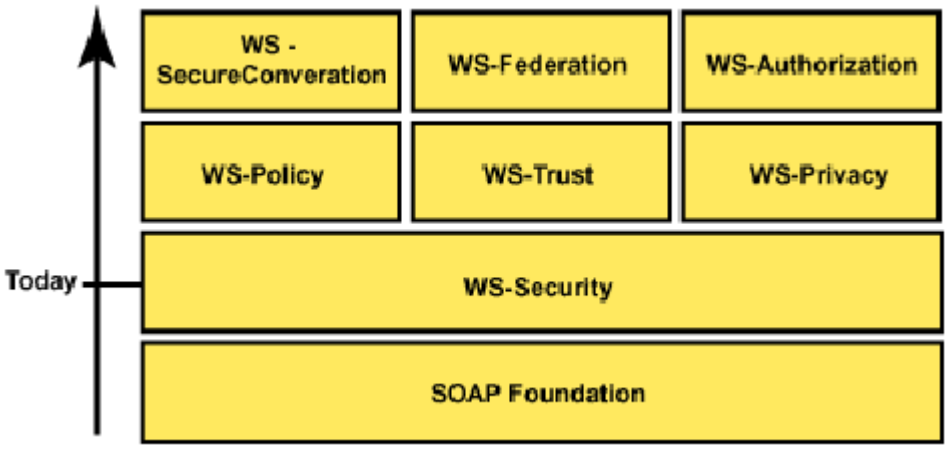
Part of the XACML language is a construct called target, a set of simplified rules with optional obligations that apply to a request. The target element is used to find applicable policy rules for the given request. Target information also provides a way to index policies for faster lookup of relevant policy rules.

Sun Microsystems [WEB23] has an open source implementation of the XACML framework for java, a project called XACML.NET [WEB24] has implemented a C# version for XACML. For Linux and Windows an XACML policy validation engine has been created by Parthenon computing [WEB25] and a beta XAXML 2.0 engine is made available on the website of AXESCON LLC [WEB26]. A prototype of implementation in the Scheme programming language [WEB27] exists although it is not intended for non-research use.

22.4.7.3 WS-*

WS-security [WEB28] is part of the WS-* standards and is a Web Services [WEB29] security specification published by IBM, Microsoft and Verisign. The specification proposes a standard set of SOAP [WEB30] extensions that can be used when building secure Web Services to implement integrity and confidentiality.

Microsoft and IBM have produced a road map [WEB31] outlining additional Web Services security specifications, which include WS-Policy, WS-Trust, WS-Privacy, WS-Secure Conversation, WS-Federation and WS-Authorization.



- WS-Security specifies how to apply XML Signatures and Encryption within a SOAP message for single-message authentication and confidentiality (origin authentication and integrity). It also specifies how to attach and refer to security tokens within a SOAP message.

- WS-SecureConversation specifies how to establish a security context and have a secure conversation with multiple messages.

- WS-Trust is the specification that defines how to request, issue, validate and exchange security tokens. A security token is a cryptographically protected claim or key (identity or authorization assertion, public, private key, etc). It also defines ways to establish and access the presence of trust relationship. The WS-security framework can support any type of security token, but tokens used over different domains should be interoperable. The following tokens are standardized and explicitly supported: username/password combinations, X.509 certificates, binary security tokens (e.g. Kerberos) and XML security tokens (e.g. SAML assertions).

- WS-Policy provides a framework to describe and publish their policies.

- WS-Federation gives a description on how to manage and broker trust relationships. It supports federated identities, attributes and pseudonyms. Mechanisms are defined for Single sign-on, sharing of attributes based on authorization and privacy policies, and integrated processing of pseudonyms.

- WS-Authorization and WS-Privacy are yet to be defined but will describe how to manage authorization data and policies, and a model for how Web Services and users state privacy preferences and statements.

While not a tool in itself, there are open source implementations available for WS-security with API that allows easy integration in existing code.

22.5 References

[AGR2002] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu. “Hippocratic databases”, in 28th Int'l Conference on Very Large Databases, Hong Kong, China, 2002.

[ALF2003] R. Alfieri, R. Cecchini, V. Ciaschini, L. dell'Agnello, A. Frohner, A. Gianoli, K. Lorentey, and F. Spataro. “Voms: An authorization system for virtual organizations,” in *proc. of the 1st European Across Grids Conference*, Santiago de Compostela, Spain, 2003.

[BIN1999] D. Bindel, M. Chew, and C. Wells. “Extended cryptographic file system,” unpublished manuscript, 1999.

[BUT2000] R. Butler, D. Engert, I. Foster, C. Kesselman, S. Tuecke, J. Volmer and V. Welch. “A National-Scale Authentication Infrastructure.” *IEEE Computer*, 2000, vol. 33, no. 12, pp. 60-66

[FEL1972] I.P. Fellegi. “On the question of statistical confidentiality.” *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 7-18, Mar. 1972.

[FOS1998] I. Foster, C. Kesselman, G. Tsudik,. And S. Tuecke. “A security Architecture for Computational Grids,” in *proc. ACM Conference on Computers and Security*, 1998, pp. 83-91.

[FU1999] K. Fu. “Group Sharing and random access in cryptographic storage file systems,” Master's thesis, Massachusetts Institute of Technology, 1999.

[GOH2003] E. Goh, H. Shacham, N. Modadugu, and D. Boneh. “SiRiUS: Securing remote untrusted storage,” in NDSS, 2003.

[GOL1998] O. Goldreich. “Secure multi-party computation”. Internet: http://www.wisdom.weizmann.ac.il/~oded/foc.html, 1998.

[KAL2003] M. Kallahalla, E. Riedel, R. Swaminathan, Q. Wang, and K. Fu." Plutus --- scalable secure file sharing on untrusted storage," in *Proc. FAST '03*, 2003, pp. 29-42.

[MIL2002] E. L. Miller, D. D. E. Long, W. E. Freeman, and B. C. Reed. "Strong security for network-attached storage," in *Proc. FAST '02*, 2002, pp. 1-13.

[NOV2001] J. Novotny, S. Tuecke, V. Welch. "An Online Credential Repository for the Grid: MyProxy," in *proc. of the Tenth International Symposium on High Performance Distributed Computing (HPDC-10)*, 2001.

[REC2003] A. Rector, J. Rogers, A. Taweel, D. Ingram, D. Kalra, J. Milan, P. Singleton, R. Gaizauskas, M. Hepple, D. Scott and R. Power. "CLEF Joining up Healthcare with Clinical and Post-Genomic Research", 2003.

[TAI2002] R. Taira, A. Bui, H. Kangarloo. "Identification of Patient Name References within Medical Documents Using Semantic Selectional Restrictions," in *Proc. AMIA Fall Symposium*, 2002, pp. 757-761.

[THO2003] M. Thompson, A. Essiari, S. Mudumbai. "Certificate-based Authorization Policy in a PKI Environment". *ACM Transactions on Information and System Security (TISSEC)*, 2003, vol. 6, no. 4, pp. 566-588,.

[TUE2002] S. Tuecke, D. Engbert, I. Foster, V. Welch, M. Thompson, L. Pearlman, and C. Kesselman. "Internet X.509 Public Key Infrastructure Proxy Certificate Profile." IETF, 2002.

[VER2004] V. Verykios, E. Bertino, I. Fovino, L. Provenza, Y. Saygin, Y. Theodoridis. "State-of-the-art in Privacy Preserving Data Mining." 2004, *ACM SIGMOD Record*, vol. 33, no. 1, pp. 50-57.

[WEB1] "The GNU Privacy Guard". Internet: http://www.gnupg.org/, Apr. 5, 2006 [Jun 14, 2006].

[WEB2] "Oracle database encryption software. The encryption wizard overview". Internet: http://www.relationalwizards.com/html/ora_encyrption.html, Mar. 7, 2006 [Jun. 14, 2006].

[WEB3] "KeepSure – SecureDB". Internet: http://www.valyd.com/Products/VDSF/databasesecurity.htm, May 26, 2006 [Jun. 14, 2006].

[WEB4] "Encrypting File System in Windows XP and Windows Server 2003". Internet: http://www.microsoft.com/technet/prodtechnol/winxppro/deploy/cryptfs.mspx, Apr. 11, 2003 [Jun. 14, 2006].

[WEB5] "EncFS Encrypted Filesystem". Internet: http://arg0.net/wiki/encfs, Apr. 1, 2006 [Jun. 14, 2006].

[WEB6] "Privacy Preserving Data Mining Bibliography". Internet: http://www.cs.umbc.edu/~kunliu1/research/privacy_review.html, May 28, 2006 [Jun. 14, 2006].

[WEB7] "De-identification". Internet: http://www.mii.ucla.edu/nlp/approach/deidentification.html, Aug. 31, 2005 [Jun. 14, 2006].

[WEB8] "The Globus Alliance". Internet: http://www.globus.org/ [Jun. 14, 2006].

[WEB9] "Globus Toolkit Version 4 Grid Security Infrastructure: A Standards Perspective". Internet: http://www.globus.org/toolkit/docs/4.0/security/key-index.html [Jun 14, 2006].

[WEB10] "Shibboleth". Internet: http://shibboleth.internet2.edu/ [Jun. 14, 2006].

[WEB11] "GridShib". Internet: http://Gridshib.globus.org/download.html, Jun. 12, 2006 [Jun. 14, 2006].

[WEB12] "Permis". Internet: http://sec.isi.salford.ac.uk/permis/, Apr. 25, 2005 [Jun. 14, 2006].

[WEB13] "Akenti: Distributed Acces Control". Internet: http://dsd.lbl.gov/security/Akenti/, Feb. 27, 2004 [Jun 14, 2006].

[WEB14] "Digital Imaging and Communications in Medicine." Internet: http://medical.nema.org/, May 25, 2006 [Jun. 14, 2006].

[WEB15] "DICOM Anonymize". Internet: http://www.codeproject.com/samples/dicom_anonymize_1.asp, Dec. 9, 2004 [Jun. 14, 2006].

[WEB16] "Innovative DICOM software for medical imaging Professionals". Internet: http://www.medisoft.gr/, Jun. 12, 2006 [Jun. 14, 2006].

[WEB17] "FP Image for Windows". Internet: http://www.fpimage.com/info.html, Mar. 10, 2000 [Jun. 14, 2006].

[WEB18] "World Wide Web Traffic Analysis Tool". Internet: http://ei.cs.vt.edu/~succeed/WebTools/, Jun. 27, 1996 [Jun. 14, 2006].

[WEB19] "Cryptography-based Prefix-preserving Anonymization". Internet: http://www-static.cc.gatech.edu/computing/Telecomm/cryptopan/, Aug. 23, 2002 [Jun. 14, 2006].

[WEB20] "OASIS Security Services TC". Internet: http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=security [Jun. 14, 2006].

[WEB21] "OpenSAML 1.1 - an Open Source SAML implementation". Internet: http://www.opensaml.org/, [Jun. 14, 2006].

[WEB22] "OASIS eXtensible Access Control Markup Language (XACML) TC". Internet: http://www.oasis-open.org/committees/xacml/ [Jun. 14, 2006].

[WEB23] "Sun's XACML implementation". Internet: http://sunxacml.sourceforge.net/, Jan. 7, 2006 [Jun. 14, 2006].

[WEB24] "XACML.NET". Internet: http://mvpos.sourceforge.net/, Mar. 27, 2005 [Jun. 14, 2006].

[WEB25] "Pathenon Computing XACML Toolkit". Internet: http://www.parthenoncomputing.com/xacml_download.html [Jun. 14, 2006].

[WEB26] "AXESCON XACML 2.0 Engine". Internet: http://axescon.com/ax2e/, Apr. 5, 2006 [Jun. 14, 2006].

[WEB27] "Margrave". Internet: http://www.cs.brown.edu/research/plt/software/margrave/download/index.html,

Oct. 27, 2005 [Jun. 14, 2006].

[WEB28] "OASIS Web Services Security (WSS) TC". Internet: http://www.oasis-open.org/committees/xacml/ [Jun. 14, 2006].

[WEB29] "Secure, Reliable, Transacted Web Services: Architecture and Composition". Internet: http://msdn.microsoft.com/webservices/?pull=/library/en-us/dnwebsrv/html/wsoverview.asp, Sep. 2003 [Jun. 14, 2006].

[WEB30] "W3C SOAP Page". Internet: http://www.w3.org/TR/soap/, Feb. 18, 2004 [Jun 14, 2006].

[WEB31] "Security in a Web Services World: A Proposed Architecture and Roadmap". Internet: http://msdn.microsoft.com/webservices/?pull=/library/en-us/dnwssecur/html/securitywhitepaper.asp, Apr. 7, 2002 [Jun 14, 2006].

[WEL2004] V. Welch, F. Siebenlist, S. Meder, L. Pearlman, and D. Chadwick. "Use of SAML for OGSA Authorization", Internet: https://forge.Gridforum.org/projects/ogsa-authz, 2004.

23 Grid Portal and Online training platforms

For ACGT to fully achieve its objectives it will need to create a community of users and a philosophy of sharing. Sharing refers to data and databases as well as tools.

In order for such a task to be achieved, a portal needs to be implemented. It will on the one hand provide a single point of entry into the ACGT space, its tools and functionality and on the other hand it will provide the technological mechanism for educating the relevant stakeholders on the use of the ACGT platform and its services.

This chapter is organised as follows. First we present the major current standards in online training. In the second part, we present two outstanding online training platforms, both from commercial and open-source parts. Finally, we investigate the state-of-the-art in the field of biomedical online training.

23.1 Relevant Projects

BRIDGES Web Portal

Delivery of the above services to the end users needs to be robust but simple since end users may be inexperienced in Grid technology and IT generally. BRIDGES therefore uses a web portal as its user interface. Various portal technologies were tested and the final choice was IBM Websphere, largely on grounds of versatility and robustness. The portal allows users to configure their individual workspaces and their settings are stored between sessions. Delivery of the applications is either through purpose-written portlets or -- where required -- by means of Java Webstart technology, which allows easy, centralised delivery of legacy java applications.

23.2 E-learning standards

In order to describe the state of the art products in the area of online training platforms we must refer to the existing standards of the domain.

Starting fromthe early 90s computer based training material was developed, and, consequently the need for standards appeared. In the United States standards have been developed by the IEEE Learning Technology Standards Committee, the IMS Global Learning Consortium, and the Aviation Industry CBT Comity; meanwhile, in Europe a corresponding standard has been developed by the Ariadne Foundation.

In January 2004, the first edition of the Sharable Content Object Reference Model (SCORM) standard appeared, developed in United States by Advanced Distributed Learning, an initiative sponsored by The Office of the Under Secretary of Defence for Personnel and Readiness. This standard includes important features of the previously mentioned standards, and adds more particularities that a Learning Management System (LMS) or a Learning Content Management System (LCMS) should respect. The second edition of SCORM was launched in December 2004, and the last version of the product is SCORM 2004 2nd Edition, Addendum 1.1.

Also there exist several research groups interested in standards for e-learning. We mention here: EifEL (European Institute for E-Learning, whose mission is to support organisations, communities and individuals in building a knowledge economy and a learning society through innovative and reflective practice, continuing professional development and the use of knowledge, information and learning technologies), CETIS (the centre for educational technology interoperability standards), from UK, and Carnegie Mellon University Learning Systems Architecture Lab (LSAL), in the US.

23.3 SCORM 2004

In the actual moment, the standard that the majority of the products refer is SCORM, since it includes the main features of the other important standards. However, it is still described, even by its developers, as the first step on the path to defining a true learning architecture. We briefly present this standard.

First, we define what a LMS is, in order to understand the description of the standard. A Learning Management System (LMS) is a software product that automates training event administration through a set of services that launches learning content and keeps track of learner progress sequences learning content. Moreover, a good LMS should respect several requirements for online education: ***Interoperability, Accessibility, Reusability, Durability, Maintainability and Adaptability***.

The SCORM standard was started with the following guidelinesin mind:

- One should be able to move courses from one LMS to another.
- One should be able to reuse content pieces across different courses.
- One should be able to sequence reusable content for branching, remediation and other tailored learning strategies.
- One should be able to search learning content libraries or media repositories across different LMS environments.

According to these guidelines, the key features that a LMS should have are: ***Sharable Content, Communication, Sequencing and Metadata***. While the first three key features are canonically linked to the above guidelines, we insist only on the concept of Metadata: by Metadata it is understood the description of the properties that an object (like a course) should have in order to be included in the system; for short it describes the properties that the data contained in the system should have.

Consequently, every SCORM-conformant LMS has a rigorously defined set of capabilities and behaviours. Different LMSs may be very different internally, but with SCORM, they interface with content in a carefully defined way. Conformant content will work with every conformant LMS. Being SCORM conformant does not affect the instructional content. The benefits of having a SCORM conformant LMS are:

- An object-based approach for developing and delivering instructional content.
- Interoperability of these objects across multiple delivery environments.
- The ability to craft sophisticated learning strategies based on the learner's mastery and progress.
- The means to package learning content and instructional strategies for import and export.
- The means to tag content so it may be found.

The Sharable Content Object Reference Model (SCORM) might be thought of as a collection of related books, like a multi-volume reference work. Each volume in the bookshelf collects

together related information and specifications, and taken together, the complete bookshelf will address the major issues and needs related to the implementation of web-based learning.

Of course, SCORM is expected to evolve over time as the state of the art advance.

We highlight below some major features of SCORM.

- **The SCORM Overview**: contains background information on the Advanced Distributed Learning (ADL) initiative, the rationale of the SCORM, and a summary of the technical specifications and guidelines encompassed by the SCORM.

- **SCORM Content Aggregation Model (CAM)** – A common method to describe the components used in a learning experience, how to package those components for exchange from system to system, how to describe those components to enable search and discovery, and how to define sequencing rules for the components.

- **SCORM Run-Time Environment (RTE)** – The RTE describes the LMS requirements for managing the run-time environment (i.e., content launch process, standardized communication between content and LMSs, and standardized data model elements used for passing information relevant to the learner's experience with the content.

- **SCORM Sequencing and Navigation (SN)** – Information and behaviors that an LMS must apply in order to present a designed learning experience. The information is expressed within Content Structure and encoded in the *organization* section of Content Packaging.

A product is said to be SCORM conformant or certified if it implements the SCORM specifications. However, only Learning Management Systems and learning content can be considered SCORM conformant or certified. There is no conformance or certification for authoring tools, organizations, individual people, etc. The official requirements for both types of products are in the Conformance Requirements documents, provided by ADL. There is one Conformance Requirements document for each edition of the SCORM.

Also, ADL has developed the SCORM Conformance Test Suite, which contains the conformance testing software, procedures and supporting documents for organizations to perform self-testing on LMSs. ADL Certification is accredited; independent testing that provides consumers of learning management systems and content with the assurance that certified products have successfully implemented SCORM. It is important to note that SCORM certification guarantees only the fact that a LMS and its content is standard, not the fact that it functions well and the content is sound. A product can be SCORM conformant, which is tested for free, but not certified; the certification costs an amount of money, which depends on SCORM version used in the tests.

23.4 Training platforms

We continue by presenting in the following sub sections two outstanding training platforms, one commercial and the other one open-source.

23.4.1 Blackboard Academic Suite

First we highlight Blackboard Academic Suite. This product consists in three different components: Blackboard Learning System, Blackboard Community System and Blackboard Content System.

The current version of Blackboard Academic Suite is Release 7.0.

Blackboard Academic Suite is tested SCORM 2004 conformant (May 4, 2006). Also, Blackboard Learning System is SCORM 1.2 certified (Jan. 9, 2004).

Blackboard Learning System offers to the user the possibility to create learning content using a variety of Web-based tools, to create particular, customized, learning paths for each of the students participating to the learning process, to facilitate the communication between the participants to the process, by implementing tools that enable synchronous and asynchronous interaction, to create and manage tests and assessments. The developers of the Blackboard Learning System focused in three major areas, following the SCORM guidelines: Instruction, Communication, and Assessment.

Blackboard Learning System supports the following features:

- **Course Management** capabilities focus on effective creation and setup of courses, as well as tools for semester-to-semester migration and archiving.
- **Content Authoring** features What You See Is What You Get editing tool that provides a rich-text editing interface similar to a word processor.
- **Adaptive Release** means instructors can create custom learning paths by determining when students can access content items, discussions, assessments, assignments or other learning activities.
- **Syllabus Builder** enables instructors to upload an existing syllabus or use the built-in creation functionality to easily design and develop their own syllabus and lesson plan.
- **Learning Units** allow instructors to create sequenced lessons and control student navigation through those lessons.
- **Course Cartridges** created by all major publishers with pre-packaged content and course materials in the Blackboard format. Course Cartridges include materials such as additional readings, updated information, multi-media and question pools.
- **Teaching and Learning Tools** enable instructors to create rich term definition lists (Glossary) as well as clearly communicate their Staff Information.
- **Personal Information Management** capabilities give students and instructors tools to better manage their work including a Calendar, asks and Blackboard Messages (course-based email).

The Personalization of the Communication System allows the information to be customized and targeted at all the levels: for users, for groups, for organizations and institutions. It is based on:

- **Role-Based Information Delivery** lets the user customize the content received through tabs, modules, channels, tools, courses and organizations based on institution roles. Users can also control and customize both the information delivered and its presentation.

- **Multi-Institution Branding and Management** facilitates separation of multiple institutions, departments or groups on one Blackboard server. Separate schools can now manage and brand their own domains.

- **Wireless/PDA** access to course and community information such as announcements, calendar items and grades is available through Blackboard Unplugged, an optional component from Blackboard Global Services.

23.4.2 Moodle

A free, open source platform, alternative to Blackboard is **Moodle**. This training platform intends to offer a learning system that permits both the development of new learning content, the integration of old resources or content developed in other frameworks, interactivity, communication between the participants at the learning process, and, nevertheless, affordability.

Moodle is **GNU free licensed**, hence all its users can access its code, and modify it according to their particular needs. Moodle is **not tested as SCORM conformant**, although it has SCORM module, which allows the user to upload any SCORM package to include in a course. The current version of Moodle is 1.5.4, but the release of Moodle 1.6 is imminent. From the design point of view, Moodle's main characteristic is modularity, in order to provide an easy modality for adding new learning content.

Regarding the learning process, Moodle proposes the constructivist approach. This means that the main activity that the learners, the students, must perform in this process is not that of learning several courses. Instead they must analyse, investigate, collaborate and share knowledge between them, build and generate new facts based on prior experience, being guided, in the same time, by educators.

Further, we discuss the features offered by Moodle. First, the Course Management is an activity performed by the teacher. It has the possibility to add the material of the course, to establish the time schedule of the course, to propose assessments to the students.

The modules of a course are:

- ***Assignment*** - Used to assign online or offline tasks; learners can submit tasks in any file format (e.g. MS Office, PDF, image, a/v etc)
- ***Chat*** - allows real-time, synchronous communication between students.
- ***Choice*** - used to create a question and a number of choices for learners; the results are posted for learners to view; this module can be used to do quick surveys on subject matter.
- ***Dialogue*** - allows for one-to-one asynchronous message exchange between instructor and learner, or learner to learner.
- ***Forums*** - Threaded discussion boards for asynchronous group exchange on shared subject matter. Participation in forums can be an integral part of the learning experience, helping students define and evolve their understanding of subject matter.

- ***Glossary*** - Create a glossary of terms used in a course. Has display format options including entry list, encyclopaedia, FAQ, dictionary style and more.
- ***Journal*** - Learners reflect record and revise ideas.
- ***Label*** - Add descriptions with images in any area of the course homepage.
- ***Lesson*** - Allows instructor to create and manage a set of linked "Pages". Each page can end with a question. The student chooses one answer from a set of answers and either goes forward, backward or stays in the same place in the lesson.
- ***Quiz*** - Create all the familiar forms of assessment including true-false, multiple choice, short answer, matching question, random questions, numerical questions, embedded answer questions with descriptive text and graphics.
- ***Resource*** - The primary tool for bringing content into a course; may be plain text, uploaded files, links to the web, Wiki or Rich Text (Moodle has built-in text editors) or a bibliography type reference.
- ***Survey*** - This module aids an instructor in making online classes more effective by offering a variety of surveys (COLLES, ATTLS), including critical incident sampling.
- ***Workshop*** - An activity for peer assessment of documents (Word, PP etc.) that students submit online. Participants can assess each other's project. Teacher makes final student assessment, and can control opening and closing period.

An important part of the system is the Learners Management. Hence, the system manages: information about learners participating to the course, the way students are grouped according to their interests and performance, information on the time schedule of each learner.

The producers of Moodle state that this product is completed by a friendly, easy to use, interface that facilitates the learners' experiences. The login to the interface is done in a usual manner; however, to participate to courses, the teacher may require an enrolment key, given to each student individually.

A very important characteristic of Moodle is that it offers multi-lingual interface support (in 34 languages), and the user can configure the time zone and language it likes to use. When participating to forums the learners are notified by e-mail of every new post. Additionally, instructors can set e-mail notification for private Dialogues. Its configurability and complex features make Moodle a viable solution both for industry and for academic learning.

23.4.3 The Sakai Project

The Sakai Project is a community source software development effort to design, build and deploy a new Collaboration and Learning Environment (CLE) for higher education. Sakai provides an application framework and associated CMS tools and components that are designed to work together. These components are for course management, and, as an augmentation of the original CMS model, they also support research collaboration.

The latest release of **Sakai** is 2.1.2 (April 2006). **Sakai** does not support SCORM yet. The main features of **Sakai** are:

- ***Announcements Tool:*** Announcements are used to inform site participants of current items of interest. Announcements can have multiple attachments, including documents and URLs.

- ***Assignments Tool:*** For courses, the Sakai Assignments Tool allows instructors to create, distribute, collect, and grade online assignments.
- ***Chat Room Tool***: The Chat Room Tool is for real-time, unstructured conversations with users who are signed on to the worksite at the same time as you are.
- ***Discussion Tool****:* The Sakai Discussion Tool allows structured conversations that are organized in categories and topics.
- ***Drop Box Tool:*** The Drop Box Tool allows instructors and students to share documents in a private folder for each student.
- ***Help Tool:*** Sakai provides an online contextual Help Tool.
- ***Membership Tool:*** The Membership Tool is a tool in one's My Workspace that allows you to join and union sites.
- ***News Tool:*** The News Tool allows a Sakai worksite to display an RSS feed. RSS is a data format that allows users to view continuously updated content from another site.
- ***Permissions and Roles****:* One can create a course or project worksite, and choose which tools (e.g., discussion, schedule, resources, etc.) the worksite will have. For each of these tools, permissions are set that allow or prevent users from seeing or performing certain tasks depending on a user's "role."
- ***Preferences:*** The Preferences Tools allows users to choose options relating to receiving emails to the site, and email notifications of new announcements and resources.
- ***Quiz & Test Tool:*** The Quiz and Test Tool allows instructors and site owners to administer online surveys, quizzes, and exams.
- ***Resources Tool:*** Resources is the most widely used tool in classes and collaborations. In Resources, you can make many kinds of material available online.
- ***Schedule Tool:*** Schedule allows instructors or worksite organizers to post items in calendar format. All Schedules on worksites one has access to, are merged in his/her My Workspace Schedule.
- ***Worksite Setup Tool:*** The Worksite Setup Tool is used to create project and course websites. It is a series of forms in steps that guide users through the process.

23.4.4 AeL

Advanced eLearning (AeL) is a commercial LMS implemented in more than 5000 K12 schools in Romania. AeL is a modern piece of software built using state of the art technologies. It is a Java based web and application server platform, database-independent multitier, portable and maintainable solution.

The main features of **AeL** are:

- Integrated solution for content management and computer assisted training
- User-friendly, intuitive interface
- Management of the organization and of the training activity
- KnowKnowledge base
- Support for reference materials
- Testing and evaluation
- Public or targeted surveys
- Instructor-lead training – virtual classroom
- Asynchronous (self-paced) study
- Integrated discussion forum
- Collaboration tools
- Monitoring
- Configurable security policies
- Easy installation, backup, recovery and administration

23.4.5 Other LMSs

WebCT is a commercial LMS used by thousands of colleges and universities in more than 70 countries worldwide.

.LRN (dotLRN – *Learn, Research, Network*) is a open-source LMS built using OpenACS (Open Architecture Community System), an advanced enterprise framework for building scalable, community-oriented web applications. .LRN is used worldwide by more than half a million users in higher education, government, non-profit and K12.

23.5 Biomedical training platforms

23.5.1 caBIG

In the following, we analyse the training platform associated with the cancer Biomedical Informatics Grid (caBIG for short). caBIG is a voluntary network connecting individuals and institutions to enable the sharing of data and tools regarding cancer research. The main goal of the network was to speed the delivery of innovative approaches for the prevention and treatment of cancer. caBIG™ is being developed under the leadership of the US's National Cancer Institute's Center for Bioinformatics.

The training platform associated with caBig is based on the idea that by consistent training and extensive documentation provided to the users of the portal it is increased its usability and it is eased its adoption. Moreover, the number of questions about the application of the presented aspects decreases, and it is ensured a common background for all the users.

The courses offered by caBig are structured in three categories, according to the way the training process is conducted:

- Web-based training sessions

- Self-paced training
- Classroom training.

The procedure for developing training modules contains 6 steps:

- Audience Analysis
- Project Scope and Purpose
- Outline and Objectives
- Training PowerPoint (Overview / Hands-On);
- Course Evaluation.

The development of the content and architecture of the courses is standardized by a Best Practices/Standard Operating Procedures. The developers of a course (training application) should add to their application the following items:

- Write down a Technical Manual that describes architecture, systems requirements, APIs, and other tools that integrate with the software being developed and their implementation.
- Write down an Installation guide, that outlines the supported configurations and technical installation instructions for the training module.
- Write down an Administration guide that describes the process for updating and maintaining, application, importing and deleting data, creating authorization for users, and users groups, on the developed training module.
- Provide Release notes, written just before the product was released, and list new features and functionalities and address known issues such as bugs and their status. Can also list appropriate documentation and websites, related to the courseware.

These items should be developed according to the templates provided on the caBig portal, ensuring that the training material provided can be evaluated in a uniform manner. Once both the developer and adopter completed the documentation required, according to the guidelines and templates provided by the portal, the training material is reviewed. Although the review does not intend to align the product to a wide spread standard as SCORM, it intends to ensure that the training material is compliant with the caBig own standards.

One can observe that in the case of caBIG the learning process is more oriented on the content of the training materials, and on the management of the interaction with users, than on the management of the students' performance, and on getting feedback for them. Indeed this strategy seems more appropriate for the purpose of the portal, since it is more addressed to users that come with a professional background, and are interested in gaining information, in learning, not in being evaluated throughout the learning process.

23.5.2 BioMed

Biomedical Online Training (BioMed for short) is a initiative driven by a in UK Consortium comprised of the Workforce Development Confederation (WDC), NHS trusts/hospitals, the

Health Protection Agency and the University of Greenwich. BioMed offers paid online courses for Employees in biotechnology and pharmaceutical industry and Healthcare scientists within the NHS using WebCT as Course Management System.

Among the problems faced by the content creators were:

- knowledge of the subject and enthusiasm for the concept, but they had little or no time available during the working day to commit to course construction;
- experience in face-to-face teaching but no experience in on-line teaching;
- some experience in using the Web, but little or no experience in preparing documents for the Web.

To tackle the problem of the lack of online course design experience, a compulsory training programme was created for authors on the Biomed project, comprised of two components:

1. Face-to-face practical training to equip authors with the practical tools necessary to publish course materials on the Internet.
2. On-line training (approximately 40hours training conducted over eight weeks) designed to
    - provide experience of being a student on an online course;
    - provide a forum for discussion;
    - facilitate course planning and development;
    - provide tips, support and guidance on how to make the online training modules interactive and motivating;
    - provide insight into what constitutes good online teaching;

Authors were required to enrol on the programme before commencing construction of the course materials, so that they would be empowered to:

- design for the on line medium, conscious of all the facilities on offer (asynchronous and synchronous communication tools; image databases; quizzes and so on);
- appreciate problems of computer access as a student;
- recognise the level of frustration that is experienced when web-links don't work;
- experience the power of interaction between students and with a tutor;
- appreciate the problem of meeting course deadlines in the face of limited time within the working environment.

23.6 Conclusions

We outline here only several major conclusions that can be extracted from the above presentation.

- Main issues in online training are Content Development, Course Management and Course Delivery.
- Online training standards are setting rules for features like Interoperability, Accessibility, Reusability, Durability, Maintainability and Adaptability.

- There are a lot of solutions (Learning Management Systems) for setting up an online training system. Among these solutions are both commercial and open-source.
- Rather commercial LMSs are standard conformant. Open-source LMSs, even if they are not formally compliant, have (or can be extended to) modules that are designed to meet the standards.
- Biomedical online training sites are less concerned with course management and delivery and more with course development.

23.7 References

[HAR2003] Patricia J. Harvey, Barry Cookson, Elizabeth, Meerabeau, Diana Muggleston, Biomedical Online Learning: The route to success, Electronic Journal of e-Learning, Vol. 1, Issue 1 (2003), 29-34.

[WEBADLa] Sharable Content Object Reference Model (SCORM) http://www.adlnet.gov/scorm/.

[WEBADLb] SCORM Certified Products http://www.adlnet.gov/scorm/certified/.

[WEBADLc] SCORM Adopters http://www.adlnet.gov/scorm/adopters/.

[WEBADV] Advanced e-Learning http://www.advancedelearning.com/.

[WEBBLA] Blackboard Academic Suite http://www.blackboard.com/products/as/.

[WEBCABa] The caBIG Training Portal https://cabig.nci.nih.gov/training/.

[WEBCABb] Developing caBIG™ Training Modules: A How-To-Guide https://cabig.nci.nih.gov/training/GuidetoTraining-FINAL.pdf/

[WEBCET] Centre For Educational Technology Interoperability Standards (CETIS) http://www.cetis.ac.uk/.

[WEBDOT] dotLRN – Learn, Research, Network http://dotlrn.org/.

[WEBEIFa] European Institute for E-Learning (EIfEL) http://www.eife-l.org/.

[WEBEIFb] E-learning standards http://www.eife-l.org/publications/standards/elearning-standard/.

[WEBGRE] Biomedical Online Training Project http://www.gre.ac.uk/biomed/.

[WEBMOO] Moodle http://moodle.org/.

[WEBSAK] The Sakai Project http://sakaiproject.org/.

[WEBWEB] WebCT (a Blackboard Company) http://www.webct.com/.

24 Web Portals

This chapter is organised as follows. First we present different approaches on web portals and compare traditional portal applications with content management systems. Secondly, we present some current standards in the development of web portals. Further, we investigate several platforms for developing web portals, including Grid-enabled portals. Finally, we investigate the state-of-the-art in the field of biomedical web portals.

24.1 Portal Technologies

24.1.1 Web Portals

Web portals are sites on the World Wide Web that typically provide personalized capabilities to their visitors. Since they are designed to deal with a great number of visitors simultaneously, the web portals use distributed applications. Different types of middleware and hardware ensure the interface between these applications and the good functioning of the portal.

The basic functionalities of a Web portal are:

- Contextualizes and frames large content sets
- Delivers personalized or customized content to audience segments or individual end users
- Manages access to published content and applications (single sign on)
- Aggregates content

The proliferation of the Web portals came together with the development of the net-browsers, in the mid 90s, when more and more companies tried to have a piece of the Internet market, by acquiring or building a Web portal. The typical services offered by Web portals are: news, interest groups, forums, chats, free email service, games etc. Important portals are: Yahoo!, AOL, Firstgov (the portal of the US government), Directgov (the UK's government portal); mini-portals are localized portals, based on local interests, that do not provide the same levels of services as major portals but they are used for collaboration of ideas, for commonly interested people (see KNET at www.silvernet.bravehost.com for example).

Most portals are written in Java, through the technology of portlets. Consequently, we will be interested in studying Java Portals and Web standards for Java portals and portlets.

24.1.2 Content management Systems

A CMS (Content Management System) is, as related to Web portals, a Web application used for managing websites and Web content (for example assisting in automating various aspects of Web publishing, like the CMSs called Wiki).

The basic functionalities of a CMS are:

- Manages the people who author content (access; rights; workflow)

- Facilitates the upholding of standards
  - A tool for applying common look and feel
- Audits the publishing environment
  - Who, what, when
- Manages the publication of content
  - Structures content for delivery

Since a large number of CMSs are available, an importat question is: “in what way one should choose an appropriate CMS for its own portal?”. This question is approached by James Robertson in the paper “How to evaluate a content management system”. The most important aspect that comes out of this paper is that a CMS should be chosen in such way that the way the steps of its life-cycle (namely Content creation, Content management, Publishing, Presentation, Contract & business) are implemented complies with the requirements of the portal. Also, the author reviews some main features that a CMS should have (for details see: http://www.steptwo.com.au/papers/kmc_evaluate/).

Most of the current CMSs are written in PHP, but there are also Java CMSs. Though, there are no standards for CMSs.

24.1.3 Web Portals vs CMSs

Web portals and CMSs have many identical features, but different teams and communities develop and maintain them. Both of them are used for building portals and Web sites. The natural question is: what are the facts that make them different (besides names)?

We can summarize the differences between a Web portal and a CMS by this definition: while CMSs are more concerned with **producing** and **managing** the content, Web portals are more concerned with **aggregating** and **delivering** the content.

However, one cannot say that CMSs are only about content production and portals only about content delivery, since both technologies have borrowed the other’s features.

24.2 Web portal standards

In order to describe the state of the art products in the area of online training platforms we must refer to the existing standards of the domain. The standards we present here mostly cover the area of Web portals.

24.2.1 Web Services for Remote Portlets

Web Services for Remote Portlets (WSRP) is a standard for Web portals to access and display portlets that are hosted on a remote server. The WSRP specification defines a web-service interface for interacting with interactive presentation-oriented web services. It has been produced through the joint efforts of the Technical Committees of the Organization for the Advancement of Structured Information Standards: Web Services for Interactive

Applications (WSIA) and Web Services for Remote Portals (WSRP). The actual version of this standard is: WSRP1.0, since August 2003. Right now, the second version of this standard WSRP2.0 is under development.

The developers of the standard motivate their work by the fact that integration of remote content and application logic into an End-User presentation has been a task requiring significant custom programming effort. It has been remarked that, typically, vendors of aggregating applications, such as a portal, write special adapters for applications and content providers to accommodate the variety of different interfaces and protocols those providers use. The goal of the WSRP specification is to enable an application designer or administrator to pick from a rich choice of compliant remote content and application providers, and integrate them with minor programming effort.

The document aims to fulfil the goal through a standard set of web service interfaces allowing integrating applications to quickly exploit new web services as they become available. The authors of the specification of the standard want to maximize the reuse of presentation-oriented, interactive web services while allowing the consuming applications to access a much richer set of standardized web services.

24.2.2 JSR 168

The second standard that we address is the Java Portlet Specification. This defines a contract between the portlet container and portlets and provides a convenient programming model for portlet developers. The Java Portlet Specification V1.0 was developed under the Java Community Process (having as members experts from leading industry companies) as JSR 168, and released in October 2003. Right now, the second version of the standard, Java Portlet Specification V2.0 or JSR286, is developed.

The Java Portlet Specification V1.0 introduces the basic portlet programming model with:

- two phases of action processing and rendering in order to support the Model-View-Controller pattern.
- portlet modes, enabling the portal to advise the portlet what task it should perform and what content it should generate
- window states, indicating the amount of portal page space that will be assigned to the content generated by the portlet
- portlet data model, allowing the portlet to store view information in the render parameters, session related information in the portlet session and per user persistent data in the portlet preferences
- a packaging format in order to group different portlets and other J2EE artefacts needed by these portlets into one portlet application which can be deployed on the portal server.

As a reference implementation of the JSR168 standard stands the portal: http://portals.apache.org/pluto/.

JSR 168 and WSRP are standards specifications that address different aspects of portlet functionality. JSR-168 is a technology specific (Java) Portlet API designed to enable interoperability between Java portlets and Java portlet containers. WSRP is a technology agnostic protocol designed to remote Portlets in a standard manner. These two are not orthogonal, but rather parallel. WSRP does not make any statements as to how the protocol should be implemented. These two specifications can be (and frequently are) used together

with JSR 168 defining a portlet and WSRP remoting that Portlet to remote containers/consumers.

24.2.3 Other Standards

The above two standards are used along side other existing or emerging web-development standards, such as:

- WSDL – Defines how abstract interfaces and their concrete realizations are defined.
- Schema – Defines how types are defined and associated with each other.
- Namespaces – Defines how XML Namespaces are declared and used.
- SOAP – Defines how to invoke web service interfaces.
- URL – Defines URI (includes URL) syntax and encoding
- XML Digital Signatures – Defines how portions of an XML document are digitally signed.
- SAML – Defines how authentication and authorization information may be exchanged.
- XACML – Defines syntax for expressing authorization rules.
- P3P – Defines how a Producer/Portlet may publish its privacy policy so that a Consumer could enforce End-User privacy preferences.
- XML Encryption – Defines how to encrypt/decrypt portions of an XML document.
- WS-Security – Defines how document level security standards apply to SOAP messages.
- RLTC – Defines syntax for expressing authorization rules.
- WS-I.org - Defining additional profiles (e.g. Security) for use of web services standards such that interoperability is maximized.
- DIME – A lightweight, binary message format that encapsulates one or more resources in a single message construct.

24.3 Web portal development platforms

We continue by presenting two outstanding web portal development platforms, one commercial and the other one open-source.

24.3.1 Liferay (portal)

Liferay Portal framework is JSR-168 compliant and runs on almost any major application server, database, and operating system, rendering over 700 deployment combinations. It provides over 50 portlets pre-bundled and more than 20 community-contributed themes available, bringing immediate usability and accelerated development potential to portal-based internet application scenarios. Liferay was built in order to have an extensive list of features that compares with most commercial portals but without the high license fees.

As important features, Liferay provides:

- **Different Themes**, that allow the user to change the look of the portal without modifying the core code
- **Different Subthemes**, permitting the customisation of the appearance of every portlet.
- **A built-in CMS**, that offers a tool for managing the information provided in the portal.
- A built-in **connector for Central Authentication Service**, Yale's single sign on engine. Liferay can also synchronize its user list between the portal and an external data source like another database or LDAP server. A default connector for Microsoft Exchange is bundled with the portal.
- Liferay was designed to be used by application service providers. This means **hosting multiple instances** of the portal (distinguished by unique URLs) on one application server and database.
- Unlike portals that come from application server vendors, Liferay is designed to be **application server agnostic** so you are not locked into a specific server. Liferay will work on lightweight servlet containers like Jetty and Tomcat, or on J2EE compliant servers like Borland ES, JBoss+Jetty/Tomcat, JOnAS+Jetty/ Tomcat, JRun, OracleAS, Orion, Pramati, RexIP, Sun JSAS, WebLogic, and WebSphere. Being a Java portal means also that this product will work on many operating systems: BSD (FreeBSD, NetBSD, OpenBSD), Linux (Fedora, Novell), Solaris, Mac OS X, and Windows.
- Liferay uses Hibernate as the ORM tool for the persistence layer which **enables pluggable databases** (DB2, Firebird, Hypersonic, InterBase, JDataStore, MySQL, Oracle, PostgreSQL, SAP, and SQL Server).
- **Portlets, CMS content, and page layouts** can all be localized to the languages wanted. The Language portlet can be easily added to any page and the end-user can select a different localization on the fly.
- Administrators can **easily manage users, organizations, locations, and roles** through a GUI interface. Access to portlets is also restricted to users based on roles. Administrators can also specify community pages so that all users who belong to a certain group see the same page.

24.3.2 Xaraya

Xaraya is an extensible, open Source web application framework written in PHP and licensed under the GNU General Public License. As important features we enumerate: this CMS is platform independent, respects standards, scalable, flexible, modular, secure, and easy to install.

In more details, Xaraya is a web application framework that allows separation of site layout, content and logic and provides a set of tools and components for building powerful web applications. The main components of the Xaraya Framework are, as described by the developers of the product:

- A slim and well defined core API
- Powerful Block Layout templating system

- Hierarchical roles based system for user and group management
- Finely tuned and robust roles based permission system
- Plug in events and authentication
- Dynamic data providing you a choice to extend data structures with or without programatic intervention
- Flexibility and custom functionality using a range of Team Xaraya and 3rd party pluggable extensions

Xaraya runs on most platforms that support PHP, and can be used with an ever growing list of relational databases including MySQL, PostgreSQL and more recently SQLite. Separate development scenarios provide support for additional databases. Xaraya's set of tools and components allows the definition of the work environment by letting the user choose how to build and customize the web site using:

- Xaraya's tools and components, without any programming, and plugging in extra functionality with dynamic data, extensions and hooks to additional site wide functions
- Xaraya development tools, powerful APIs and reusable code to develop complex applications in a rapid development environment
- A mixture of both the above methods

These components can be used to build web applications or systems such as:

- a simple static document website
- a dynamic database driven community site
- a content management system or news publishing site
- personal blog
- a company intranet
- portals
- a specialized industry site with custom built Xaraya applications
- an enterprise level multi-site and multi-language deployment

24.4 Grid-enabled portal development platforms

24.4.1 GridSphere (portal)

The starting point of the developers of GridSphere was the desire to build a Portal, similar with the usual Web portals such as Yahoo or Amazon, that will be able deliver the benefits of Grid computing to virtual communities of researchers and scientists, providing customizable, easy-to-use, singular access points to Grids.

GridSphere, is an open-source JSR-168 compliant portal framework that is ready to run with a suite of tutorial and example portlets. One of the key elements in GridSphere is the ability for site administrators and individual users to be able to dynamically configure the content they choose to see based upon their application needs. To make content dynamic, individual components in the portal are generally engineered independently from one another. One of the key benefits to using GridSphere comes from the additional package, GridPortlets, which

provides many of the portlets needed to produce a production Grid-portal. The main benefits are, thus, the fact that the portal is JSR-168 compliant, it uses GridPortlets and Portlet Services framework. Also it has a large user-base and support community.

The technical features listed by the developers of GridSphere consists in:

- Portlet API implementation compatible with IBM's WebSphere 4.2.
- Support for the easy development and integration of "third-party portlets"
- Higher-level model for building complex portlets using visual beans and the GridSphere User Interface tag library.
- Use of Style Sheets and User Interface tags in order to allow GridSphere to be “themable”
- Flexible XML based portal presentation description can be easily modified to create customized portal layouts.
- Built-in support for Role Based Access Control separating users into guests, users, admins and super users.
- Sophisticated portlet service model that provides functionality that can be reused across multiple portlets.
- Persistence of data provided using Hibernate supports most major databases including MySQL, Postgres, DB2, HsqlDB, etc, as in the above presented case of Liferay.
- Integrated Junit/Cactus unit tests for server side testing of portlet services including the generation of test reports.
- GridSphere core portlets offer base functionality including login, logout, user and access control management.
- Full localization support in the Portlet API implementation and GridSphere core portlets supporting several languages.
- Together with the Core services (Portlet Manager Service, that provides lifecycle methods to allow portlets to be installed, removed, initialized and destroyed by authorized users, Login Service, that allows a User to be retrieved from a username and password, User Manager Service, add/remove user accounts, edit user profiles, access control service, add/remove user groups, add/remove user roles), an important role is played by the Grid Services such as:
- Credential Manager Service, used to add/remove allowed User Credentials and configure use of Credential Retrieval Service
- Job Manager Service, used for listing, starting, migrating, stopping jobs.
- Job Monitoring Service, used to specify what to monitor for any given job and archive related information.
- File Transfer Service, useful for managing and scheduling file transfers.
- Data Manager Service, used to access to data replica catalogues, and describe data with meta-data.
- Notification Service, used to Define events to be notified about, and to specify how to be notified about those events.

While the Core services are implemented via core portlets, the Grid services are implemented via Grid portlets:

- Credential Administrative Portlets. By these portlets it can be specified what credentials are permitted for use, the mappings between credential subjects and user accounts, as well as mappings to particular resources. Also, admins can view active credentials and their usage online.

- Credential User Portlets. By these portlets users may request new credential mappings to their accounts, and may retrieve and refresh credentials for later use.

- Resource Management Portlets. These are used to specify and describe Grid resources, as tools for discovering resources on the Grid, and as tools for tracking requests made to site admins for configuring or updating resources with software, etc.

24.4.2 Grace (CMS)

GRACE (Grid seArch & Categorization Engine) is a content management system designed to enable research communities to get organized around their specific common interests, and share their computational storage and knowledge resources according to their specific information needs.

The developers of GRACE state that the product is based on the principle that a content management system must not replace or unnecessarily enlarge the customer's existing resources, but rather allow the customer to maximize their use. To obtain this, two major advantages were implemented: integration of the existing content sources and Grid technology.

GRACE also implements an innovative approach to the integration of multiple content sources: it systematically harvests relevant information from these documents, applying very strong natural language processing methods in order to re-index them into Knowledge Domains. Knowledge Domain is not only a complete virtualization of multiple relevant content sources, but also incorporates the underlying semantics encapsulated in related ontologies. These ontologies are used to: query the content sources, and than index them by associating them with the key terminology; to extract from the extensive content sources only the portion that is indeed relevant for a particular interest; the ontologies are further used for querying, just like a normal index, browsing, and presentation of the search results. Moreover, it is performed a constant update of the Knowledge Domains with new and relevant information.

Also, the utilization of the strong natural language processing methods allows GRACE to offer unprecedented Information Retrieval functionalities to the knowledge workers. GRACE provides multilingual abilities for several languages, as: English, German, Italian, but other languages can be added by integrating suitable lexical databases.

GRACE utilizes the Data Grid (a special case of Grids that assume that data consumed by these computations must be shared across an organization and that it is thus more efficient to keep these data stored in a central location, accessible to various front end applications targeting various organizational information needs) in order to make the Knowledge Domains securely accessible throughout an organization, regardless of the geographical location or the strength of the locally available resources.

The developers of GRACE, state that their product is the first working system that allows integration of the unstructured, textual information.

24.5 Biomedical web portals

24.5.1 caBIG

The cancer Biomedical Informatics Grid, or caBIG, is a Grid connecting individuals and institutions to enable the sharing of data and tools, creating a World Wide Web of cancer research. The goal is to speed the delivery of innovative approaches for the prevention and treatment of cancer. The infrastructure and tools created by caBIG also have broad utility outside the cancer community. caBIG is being developed under the leadership of the National Cancer Institute's Center for Bioinformatics.

The current test bed architecture of caBIG is caGrid. The software embodiment and corresponding documentation of this architecture constitute the caGrid 0.5 release. The caGrid 0.5 software release contains tools for creating and deploying caBIG-compliant Grid services to the caGrid 0.5 infrastructure. The software infrastructure is designed to satisfy the use case requirements from the various caBIG Domain Workspaces (such as: Clinical Trial Management Systems, In Vivo Imaging, Integrative Cancer Research, Tissue Banks and Pathology Tools). caGrid 0.5 is providing the necessary infrastructure for caBIG applications to leverage the following Grid infrastructure capabilities:

- Indexing and Registry Services
- Metadata Management
- Common Data Elements
- Controlled Vocabulary Semantics
- XML Schema Management
- Security Services
- Discovery and Invocation
- Data Service Toolkit
- Analytical Service Toolkit

24.5.2 Telescience

In 1999, web-based Telemicroscopy was released and researchers were able to effectively use the remote interface to acquire data. It became clear that for a complete remote research scenario, the ability to remotely acquire data must be closely coupled to data computation and storage resources. The Telescience Project was developed to address this issue.

Telescience, provides a Grid-based architecture to combine the use of Telemicroscopy with tools for:

- Parallel distributed computation,
- Distributed data management and archival, and
- Interactive integrated visualization tools

to provide an end-to-end solution for high throughput microscopy.

The Telescience Project merges technologies for:

- Remote control,
- Grid computing, and
- Federated digital libraries of multi-scale, cell-structure data.

The objective of the Telescience Project is to increase the throughput of data acquisition and processing and ultimately improve the accuracy of the final data product.

The Telescience Portal represents the interface for the Telescience Project. It consolidates access for controlling instruments remotely, managing data, and controlling batch jobs with a single login and password. The Portal walks the user through the complex process of remote data acquisition via Telemicroscopy; Globus-enabled parallel tomographic reconstruction; advanced visualization, segmentation, and data processing tools; and transparent deposition of data products into federated libraries of cellular structure. Key features of the Portal include personalized user information, collaboration tools such as chat and shared white boards, and automatic storage of data and job tracking tools.

There are currently several portals supported under the Telescience project:

- NCMIR Multi-Scale Imaging Portal (the original Telescience-based Portal) http://swilken.ucsd.edu:8080/Gridsphere/Gridsphere
- NIEHS Environmental Health Science Data Resource Portal http://balata.ucsd.edu:8080/Gridsphere/Gridsphere
- Biomedical Informatics Research Network Portal https://portal.nbirn.net/BIRN/
- Branfman Family Foundation Collaboration Portal http://pebblebeach.ucsd.edu:8080/Gridsphere/Gridsphere
- NCI National Brain Cancer Model Collaborative Network (nBCMn) Portal http://birkdale.ucsd.edu:8080/Gridsphere

All these implementations are based on Gridsphere technology.

24.6 Initial Conclusions

We outline here only several major conclusions that can be extracted from the above presentation.

- Web portals and especially the ones using opensource technologies are based on standard Java portlets.
- Sometimes web portals are replaced by Content Management Systems (CMSs).
- Web portals and CMSs have many common features, but, basically CMS are more concerned with the production of content, while Web portals are more concerned with the delivery of content.

24.7 References

[WEBOPEa] Opensource Content Management Systems basically written in PHP and MySQL HTTP://WWW.OPENSOURCECMS.COM/

[WEBJAVa] Opensource Content Management Systems written in Java HTTP://JAVA-SOURCE.NET/OPEN-SOURCE/CONTENT-MANAGMENT-SYSTEMS

[WEBJAVb] Opensource Web Portals written in Java HTTP://JAVA-SOURCE.NET/OPEN-SOURCE/PORTALS

[WEBSTE] J. Robertson, How to evaluate a content management system HTTP://WWW.STEPTWO.COM.AU/PAPERS/KMC_EVALUATE/

[WEBWEL] L. Welchman, CMS vs Portal – Which one do you need, HTTP://WWW.WELCHMANCONSULTING.COM/PRESENTATIONS/CMS_PORTAL.PPT

[WEBOPEb] Opensource Grid-enabled Web Portals HTTP://WWW.OPENGRIDPORTALS.ORG/

[WEBGRI] Gridsphere Webpage HTTP://WWW.GRIDSPHERE.ORG/

[WEBGRA] Grace Webpage HTTP://WWW.GRACE-IST.ORG/

[WEBLIF] Liferay Webpage HTTP://WWW.LIFERAY.COM/

[WEBXAR] Xaraya Webpage HTTP://WWW.XARAYA.COM/

[WEBCAB] The caBIG Portal HTTPS://CABIG.NCI.NIH.GOV/

[WEBTEL] The Telescience Project HTTP://TELESCIENCE.UCSD.EDU/

PART 3

Appendices

25 Appendix 1 - Conduct of a clinical trial

For practical reasons the steps that are neede to take to to conduct a clinical trial are described in the following part. This description is mainly based on a Guide used at the Great Ormond Street Hospital in London This algorithm is based on the UK legislation, but is transferable to other European countries and should be checked by WP 10. The following figure outlines seven steps that have to be taken.

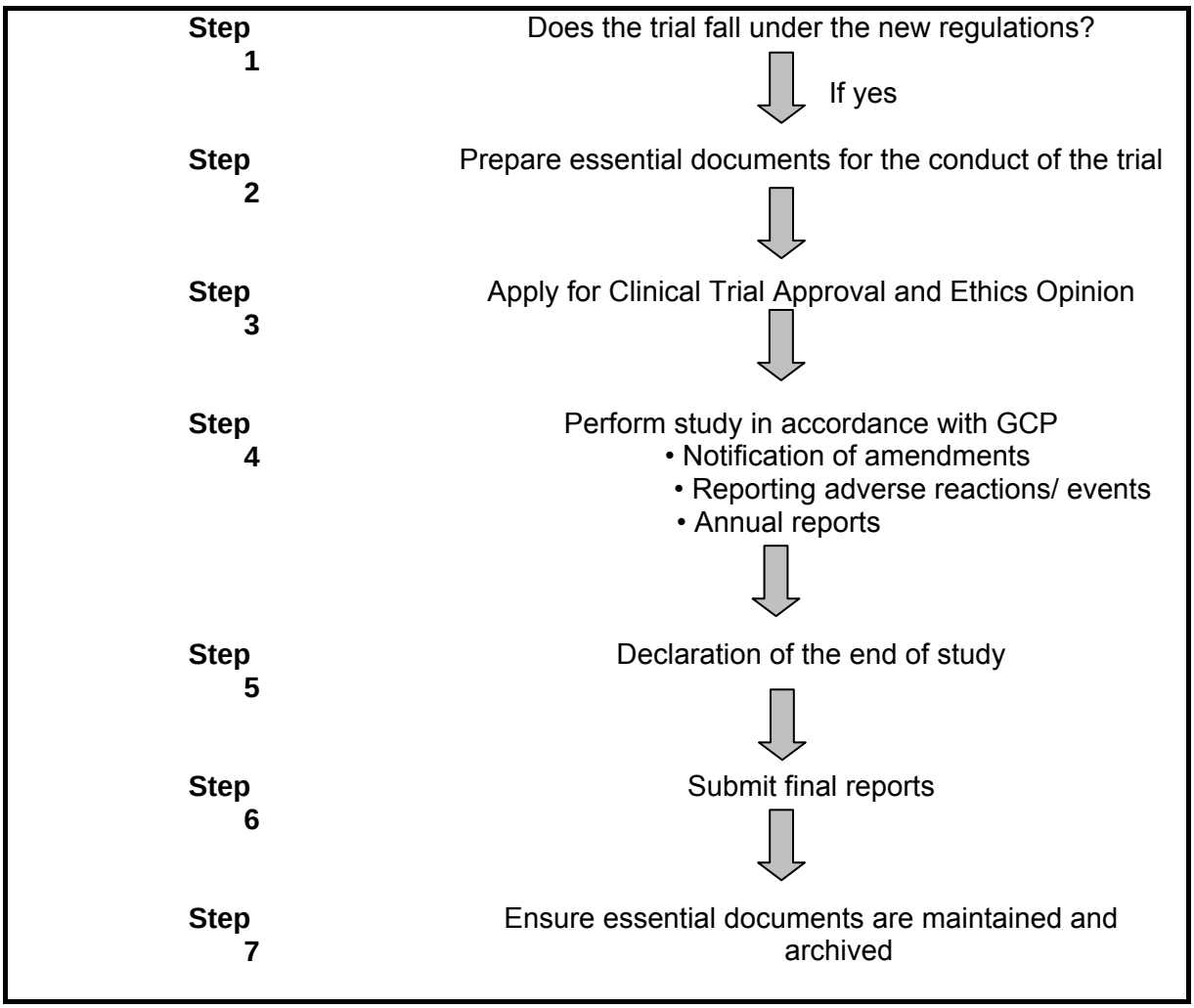


Figure 43: Outline of steps for the conduct of clinical trials

26 Appendix 2 – Internal Inquiry on Systems and Tools

As ACGT partners come from multiple domain areas, the project has launched an *internal inquiry* among them with the intention to benefit from a synergy of expertise which would allow dividing and conquering the problem effectively.

The inquiry was divided in three different documents structured as follows.

- A Word document containing the "skeleton" of the inquiry with questions and references to the rest of the documents comprising the inquiry.
- An Excel document containing questions involving listing several items.
- A Word document containing a glossary of terms.

The main questions been asked were:

- Which *scientific tools* do you use and might be of importance for ACGT
- Which *development tools* do you use and might be of importance for ACGT
- Which tools or software have *to be created* and might be of importance for ACGT
- Which *standards* may be adopted by ACGT
- Which scenarios may be supported by ACGT
- Provide names of clinicians and molecular biologists
- Glossary of terms

26.1 Inquiry outcome

The inquiry started by a proposal of an initial draft which consisted on an Excel document containing the main questions about available tools and desired functionality. Later on, it was extended with a glossary of terms and questions relating to adoption of standards, definition of clinical scenarios and requirements.

As part of the process for defining the inquiry, the questionnaires were reviewed and comments were given by partners. The inquiry was launched at the beginning of May 2006, with the finally version of the questionnaires. Since then all collected answers to the questionnaire have been curated and added to the original forms, creating a well organized and non-redundant view of the answers. Finally, a report has been written presenting and providing an initial analysis of results.

26.2 Enquiry results

At this point we have a detailed enumeration of databases and tools that are of interest for ACGT aims. They have been organized by domains and priority in terms of importance to ACGT has been assigned to them. High priority has been given to services and data sources

which should be part of the minimal system. First, we present a summary followed by two tables containing the full information about the collected list of tools and databases.

It is important to mention that this summary represents a set of recommendations to develop a functional set of data repositories and services. This does not imply product development , since some of them are commercial products, but rather implement the required data access or tool invocation functionality.

26.2.1 Result summary

Databases:

- **High**: BIC, OMIM, BASE, ArrayExpress, GEO, EMBL Bank, EnsEMBL, GenBank, UniGene, Swissprot, UniProt, PDB, PAM, BLOSUM, GO, GOA, KEGG, PubMed, BEA, Ensembl, UniProt, KEGG, Pubmed.
- **Medium**: Oncomine, SMD, PIR, CATH, PDBsum, FSSP, DSSP, HSSP, IntAct, AMIGO, Pfam, Prosite, NCBI databases, CATH, PDBsum, phenotypes, CAS, SMART, BIND, DIP, MINT, PIM,.
- **Low**: CleanEx, DDBJ.

General bioinformatics tools:

- **High:** Conversion / translation**,** Parsing / extraction, Retrieval, LIMS.
- **Medium:**
- **Low:**

Genomics:

- **High:** BLAST, WU-Blast2, PSI-BLAST, FASTA, EMBOSS package.
- **Medium**: FASS, DNADIST, jdotter, T-Coffee, ClustalW & ClustalX, DAVID, SAAM II, TransFind, MEME/MAST, Mutational behaviour, HMM, HMMER, HMMSearch , phylogenetic studies, PHD, Hot Spots, PMUT.
- **Low**: Contigs and assembling, Sequence characterization, Gene identification, Vector NTI, Mutational behaviour, Functional residues prediction, Motif database searching, Sequence / structure comparison, Comparative analysis of 3D structures.

Transcriptomics (DNA Microarray tools)

- **High:** GeneChip Operating System (GCOS), GenePix, Scanalyze2, Engene, Expression Profiler, PreP, Xcluster, Cluster & TreeView, R - Bioconductor packages.
- **Medium**: Oligo, Primer3, BioMine, GEPAS, J-Express, MAExplorer, Microtracker, Partek Software Suites, Association rule discovering, LitheMiner.
- **Low**: CSIRO Spot, spotSegmentation, Acuity, AMADA / AMIADA, ArrayStat, BRB Array Tools, DNA-array analysis tools, DNA-Chip Analyzer, ExpressionSieve, GeneLinker, GeneMaths, GeneSight, GeneSpring GX, Genesis, GeneSense, Soochika, TIGR MIDAS, X-Miner, Xpression, Qfirst, GenMapp, ImaGene, MeV,

MIDAS, PathwayStudio, Ingenuity Pathway Analysis (IPA), Decision site, TIGR SpotFinder, SAGE, MineBioText, MineGene, ArrayUnlock.

Data mediation / integration

- **High:** FatiGO, PubMed, DAS systems.
- **Medium**: Protégé 3.1, bioBroker.
- **Low**: FACT++/Pellet, Jena Framework, OWL, BioMart, Annotations (FunCUT).

Proteomics

- **High:**
- **Medium:** Melanie & ImageMaster, Decyder (Amershan Pharmacia), Phoretix 2D (Phoretix International), Gellab, Kepler; Z3; GD-Impressionist, Progenesis PG600.
- **Low:**

26.2.2 Data sources

Databases of importance for ACGT				
Name	Description	more info	Lincese / Price	Priority (H/M/L)
Medical Databases				
BIC	Breast cancer Information Core (public database)	http://research.nhgri.nih.gov/bic/	Free for academic	High
OMIM	Online Mendelian Inheritance in Man, a database of human genes and genetic disorders	http://www.ncbi.nlm.nih.gov/omim/	Free for research	High
Gene Expression Databases				
BASE	A LIMS system (ref: Lao H. Saal, Carl Troein, Johan Vallon-Christersson, Sofia Gruvberger, Åke Borg and Carsten Peterson BioArray Software Environment: A Platform for Comprehensive Management and Analysis of Microarray Data)	http://base.thep.lu.se/	Free	High
ArrayExpress	MIAME compliant repository of published microarray datasets (EBI)	http://www.ebi.ac.uk/arrayexpress/ http://www.ebi.ac.uk/miamexpress	OS	High
GEO -Gene Expression Omnibus.	A gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval. (Microarray database with good search engine and export of data)	http://www.ncbi.nlm.nih.gov/geo/	OS	High
Oncomine	Microarray Database Oncomine (includes some tools for profiling)	www.oncomine.org	OS	Medium
CleanEx	Comparative database of published microarray datasets.	http://www.cleanex.isb-sib.ch/	Free	Low

SMD	Stanford Microarray Database (Database of Stanford arrays with export of data and some tools for filtering and first analysis)	http://genome-www5.stanford.edu/	OS	Medium
Nucleotide Sequence Databases				
EMBL Bank	European Nucleotide Sequence Database	www.ebi.ac.uk/embl	Free	High
EnsEMBL	Integrated nucleotide sequence knowledge base	http://www.ensembl.org/	Free	High
GenBank	American Nucleotide Sequence Database	www.ncbi.nih.gov/ www.ncbi.nlm.nih.gov/Genbank/	Free	High
UniGene	Database of clusters of GenBank sequences.	http://www.ncbi.nih.gov/	Free	High
DDBJ	DNA Data Bank of Japan	http://sakura.ddbj.nig.ac.jp/		Low
Protein Databases				
Swissprot	Protein knowledge base	http://www.expasy.org/sprot/	Free	High
UniProt	Universal protein resource: Consolided DB from : Swissprot + TrEMBL + REMTrEBML + PIR).	http://www.uniprot.org	free to copy, distribute, ...	High
PIR	Protein information resource	http://pir.georgetown.edu/	Free	Medium
3D Structure databases				
PDB	Protein data bank	http://www.rcsb.org/pdb/	Free	High
CATH	Protein structure classification. CATH is a hierarchical classification of protein domain structures, which clusters proteins at four major levels, Class(C), Architecture(A), Topology(T) and Homologous superfamily (H).	http://cathwww.biochem.ucl.ac.uk/		Medium
PDBsum	Putative protein-protein binding sites, ligand binding sites, and protein-DNA binding sites by homology with those observed in crystallized protein structures.	http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/	freely available	Medium

FSSP	Fold classification based on structure-structure assignments	http://www.ebi.ac.uk/dali		Medium
DSSP	secondary structure assignments for all PDB-protein entries (it is also a programm)			Medium
HSSP	DB of homology-derived secondary structure of proteins	http://swift.cmbi.kun.nl/gv/hssp/		Medium
IntAct	Protein interaction data	http://www.ebi.ac.uk/intact/	freely available	Medium
Distance Matrices (Distance matrices are needed in all programs that perform sequence comparison)				
PAM	Point Accepted Mutation matrices (from PAM10 to PAM450)		free	High
BLOSUM	Block alignment derived substitution matrices (from Blosum 30 to Blosum90)		free	High
Ontology Databases				
GO: Gene Ontology	Function, Biological process, and Cellular component	http://www.geneontology.org/	free	High
GOA	Gene Ontology Annotation @ EBI, provides association between GO terms and genes	http://www.ebi.ac.uk/GOA/	Open access	High
AMIGO	Gene Ontology database	http://www.godatabase.org/cgi-bin/amigo/go.cgi	OS	Medium
Pathway Databases				
KEGG	Kyoto Encyclopedia of Genes and Genomes: pathways map	http://www.genome.ad.jp/kegg/	Licenses for non-academic users	High
EBI Databases repositorie	Public databases and tools	http://www.ebi.ac.uk/	Free	High
Motif Databases				
Pfam.	Protein Families and domains database (includes multiple sequence alignments and HMM models	http://www.sanger.ac.uk/Software/Pfam/		Medium
Prosite, BLOCKs, PRODOM, PRINTS..	Sequence motifs databases; protein domains, etc	www.expasy.org/prosite/ http://protein.toulouse.inra.fr/prodom/current/html/home.php		Medium
Scientific literature				

PubMed	Bibliographic references (including MeSH terms) . PubMed is a service of the U.S. NLM that includes over 16 million citations from MEDLINE and other life science journals for biomedical articles back to the 1950s.	http://www.ncbi.nlm.nih.gov/PubMed/	Public domain access	High
NCBI databases	Public databases and tools	http://www.ncbi.nlm.nih.gov/	OS	Medium
BEA	Biovista	http://www.biovista.com	Commercial	High
DAS servers (data integration)				
Ensembl	system which produces and maintains automatic annotation on selected eukaryotic genomes	http://www.ensembl.org	free access	High
UniProt	Protein features annotations	http://www.ebi.ac.uk/uniprot-das/ (DAS server)	free access	High
KEGG	Pathway maps	http://www.genome.jp/kegg/soap/ (Java)	Licenses for non-academic users	High
GO: Gene Ontology	Function, Biological process, and Cellular component	http://www.geneontology.org/GO.tools.shtml (different annotation tools)	freely available to all the public	High
Pubmed	Bibliographic references (including MeSH terms)	http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html	Public domain information (NLM)	High
CATH	Annotates PDB structures with CATH structural domains	http://www.biochem.ucl.ac.uk/bsm/cath/ (DAS)		Medium
PDBsum	Putative protein-protein binding sites, ligand binding sites, and protein-DNA binding sites by homology with those observed in crystallized protein structures.	http://www.ebi.ac.uk/das-srv/proteindas/das/sasprot/ (DAS)	freely available	Medium
phenotypes	Phenotypes associated directly or via orthologues or protein families. Use the Ensembl,Gene_ID databases.	http://www.ebi.ac.uk/das-srv/genedas/das/phenotypes/ (DAS)		Medium
Catalytic Site Atlas (CAS)	Manually curated collection of catalytic sites (and predicted by homology) described in the literature	http://www.ebi.ac.uk/das-srv/proteindas/das/csalit/ and http://www.ebi.ac.uk/das-srv/proteindas/das/csaextended/ (DAS)		Medium

SMART	Domain annotations for Uniprot/Ensembl	http://smart.embl.de/smart/das/smart/ (DAS)	http://smart.embl-heidelberg.de/	Medium
BIND, DIP, MINT, PIM, etc...	Protein interactions	http://www.blueprint.org/bind/bind_relateddatabases.html		Medium

26.2.3 Services

Scientific tools for bioinformatics and biomedicine relevant to ACGT				
Area / tool	Description	More info	Lincese / Price	Priority (H/M/L)
General bioinformatics tools				
Conversion / translation services	change format (without loss of data)	i.e. converts an alignment in Clustalw format into Phylip format / or an aminoacid FASTA sequence into an AA sequence		High
Parsing / extraction data services	extract partial information	i.e. obtain the AA sequence from a structural entry in PDB / or get the best hits from a Blast report using the E-Value as threshold		High
Retrieval services	Retrieves an entry for a given DB	i.e. records from EMBL, GenBank, Swiss-Prot / or partial information such as the 3D coordinates from a PDB entry		High
Data translation services	i.e. translate a nucleotide sequence in all its 6 ORFs (coding as AA sequence)			High
LIMS	Laboratory Information management system (for gene expression data)	ArrayHub (integromics.com)	Commercial	High
Genomics				
Functional genomics				
Contigs and assembling	Software for editing and assembly of sequence chromatograms	(i.e Vector NTI)		Low

Sequence characterization	Software for determining different sequence properties, such as: physicochemical; CG content, hidrophobicity, etc			Low
Gene identification	An initio gene prediction tools	(i.e. GeneID at genome.imim.es)		Low
BLAST	The Blast family of programs for "similarity searching" are focused in finding regions of sequence similarity quickly, with minimum loss of sensitivity	http://130.14.29.110/BLAST/	Free	High
WU-Blast2	The Washington University Blast tool (version 2.0). This is the basic tool which includes statistics	http://www.ebi.ac.uk/blast2/	Free academy	High
PSI-BLAST	Position specific iterative BLAST (PSI-BLAST) refers to a feature of BLAST 2.0 in which a profile (or position specific scoring matrix, PSSM) is constructed	http://www.incogen.com/public_documents/vibe/details/NcbiBlastp.html	Free academic	High
FASTA	FASTA family of programs for biological sequence comparison programs for searching protein and DNA sequence databases. These programs are equivalent to Blast family using a different approach	http://fasta.bioch.virginia.edu	Free	High
EMBOSS	Pairwise global alignment using the Needleman-Wunsch algorithm.	http://www.compbio.ox.ac.uk/analysis_tools/EMBOSS.shtml	Free Open Source	High
FASS	Predicting Functional Residues in Protein Sequence Alignments	http://somosierra.cnb.uam.es/Servers/treedet/	free	Medium
DNADIST	Computes a distance matrix from a set of nucleotide sequences, under different models of nucleotide substitution.	(http://evolution.genetics.washington.edu/phylip.html) Phylips package		Medium
jdotter	Dotplots	http://athena.bioc.uvic.ca/techDoc/jdotter/		Medium
T-Coffee	Multiple sequence alignment combination	http://igs-server.cnrs-mrs.fr/~cnotred/Packages/T-COFFEE_distribution_Version_4.45.tar.gz	Free	Medium

ClustalW & ClustalX	Alignment / similarity search tools. CLUSTAL W: perform a progressive MSA through sequence weighting, position-specific gap penalties and weight matrix choice.	http://www2.ebi.ac.uk/clustalw/	Free	Medium
DAVID	Data annotation software. (ref: PMID: 12734009)	http://david.niaid.nih.gov/	Free	Medium
SAAM II pharmacokinetics	Simulation, Analysis and Modelling for Kinetic Analysis	http://depts.washington.edu/saam2/		Medium
Vector NTI	Desktop application for DNA and protein sequence analysis. Support annotation, assembly, blast searching, translation, primer design, in silico cloning, and a number of other common tools (includes graphical interface)	http://www.invitrogen.com/content.cfm?pageid=10071	Commercial	Low
TransFind	Pattern discovering in protein sequences	http://www.genomicdiscoverytools.com/	Commercial	Medium
MEME/MAST	Pattern discovering in protein sequences	www.inab.org/MOWServ	Free upon request	Medium
Mutational behaviour	Automatic method for predicting functionally important residues (Mutational Behaviour) in protein sequence alignments (longer than 50 residues and at least 15 sequences), in the FASTA text format, as input.	pdg.cnb.uam.es	Free upon request	Low
Functional residues prediction	Automatic method for predicting functionally important residues using the concept of Relative Entropy in protein sequence alignments	www.inab.org/MOWServ	Free upon request	Low
Hidden Markov Models (HMM) for sequence profiling	Constructs a Hidden Markov Model from a multiple sequence alignment.	SAM software (sequence alignment and modelling system) http://www.cse.ucsc.edu/compbio/sam.html	Free	Medium
HMMER	Searches a HMM profile database (Pfam) with a query sequence (identify those HMM profiles embedded in the sequence)	http://hmmer.wustl.edu/	Free	Medium
HMMSearch	Searches a set of sequences with a HMM profile (identify those sequences that	http://pfam.cgb.ki.se/hmmsearch.shtml	Free	Medium

	contains the profile)			
Motif database searching	Analyzes a DNA sequence for putative specific factors (i.e. binding sites) from specific databases (i.e. transfac)	i.e. factor binding sites from Transfac or Jaspar		Low
phylogenetic studies	Phylip Package (Maximum likelihood, parsimony, Fitch-Margolias, Neighbor join UPGMA,	http://evolution.genetics.washington.edu/phylip.html	free	Medium
Structural genomics				
PHD	Performs secondary structure and accessibility predictions using PHD	http://mmb.pcb.ub.es/	Free academy	Medium
Hot Spots	Predicts sequence positions (Hot Spots) that would produce pathological behaviour when mutated. Trained with human pathological mutations	http://mmb.pcb.ub.es/	Free academy	Medium
PMUT	Annotation and prediction of pathological mutations (given a mutation at a specific location in a protein sequence, estimate whether it can be pathological (it can lead to disease for the carrier) or non-pathological / neutral (no effect on the carrier's health)	http://mmb2.pcb.ub.es:8080/PMut/	Free academy	Medium
Sequence / structure comparison	Per residue reliability for alignments between query sequences and sequences of known structure (ie those from the PDB)	pdg.cnb.uam.es		Low
Comparative analysis of 3D structures	Protein fold classification, active site, residues maps (using 3D databases: SCOP, CATH, DALI, etc.)	SCOP: http://scop.mrc-lmb.cam.ac.uk/scop. CATH: http://www.biochem.ucl.ac.uk/bsm/cath. DALI/FSSP: http://www2.embl-ebi.ac.uk/dali.		Low
Transcriptomics (DNA Microarray tools)				
Experiment design				

Oligo	Primer & Probe Design Tools	www.oligo.net	Commercial	Medium
Primer3	Primer & Probe Design Tools. (ref: Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers.	http://frodo.wi.mit.edu/primer3/primer3_code.html	Free	Medium
Data acquisition software				
CSIRO Spot	Microarray/Macroarray Image Processing Tools	http://experimental.act.cmis.csiro.au/Spot/index.php	Commercial	Low
GeneChip Operating System (GCOS)	Affymetrix image analysis tool	(W) http://www.affymetrix.com/	Commercial (free)	High
GenePix	Microarray image segmentation and analysis tool	http://www.moleculardevices.com/	Commercial	High
Scanalyze2	Microarray/Macroarray Image Processing Tools	http://rana.lbl.gov/index.htm	Free academy	High
spotSegmentation	Microarray/Macroarray Image Processing Tools	http://www.stat.washington.edu/fraley	GPL	Low
Acuity	Axon Instruments Inc.	(MS.W) http://www.axon.com	Commercial	Low
AMADA / AMIADA	Dept. Biol. Univ. Ottawa	(MS.W) http://aix1.uottawa.ca/~xxia/software/amiada.htm	Free	Low
ArrayStat	Imaging Research Inc.	(MS.W) http://www.imagingresearch.com	Commercial	Low
BioMine	Gene Network Science	(MS.W) http://www.gnsbiotech.com	Academic license	Medium
BRB Array Tools	Mol. Stat. and Bioinf. Section, Biometric Res. Branch, NCI	(MS.W) http://linus.nci.nih.gov/BRB-ArrayTools.html	Free academic & non-profit	Low
DNA-array analysis tools	National Spanish Cancer Center (CNIO)	(Web) http://bioinfo.cnio.es/dnarray/analysis	free	Low
DNA-Chip Analyzer	Wong Lab. Dept. of Stat., Hard-vard Univ.	(MS.W) http://www.dchip.org	Academic license	Low
Engene	Comp. Archit. Dept., Univ. of Malaga.	(Web) http://www.engene.cnb.uam.es; chirimoyo.ac.uma.es/engenet	free upon request	High
Expression Profiler	European Bioinf. Instit. (EBI)	(Web) http://ep.ebi.ac.uk	free	High

ExpressionSieve	BioSieve	(Java) http://www.biosieve.com	Academic license	Low
GeneLinker, GeneLinker Platinum	Mol. Mining Corp.	(MS.W) http://microarray.genelinker.com	Academic license	Low
GeneMaths	Applied Maths	(MS.W) http://www.applied-maths.com	Commercial	Low
GeneSight	BioDiscovery	(MS.W-Linux-Mac) http://www.biodiscovery.com	Commercial	Low
GeneSpring GX	Agilent (formerly Silicon Genetics) General purpose microarray mining tool.	(MS.W-Linux-Mac)http://www.chem.agilent.com	Commercial	Low
Genesis	Bioinf. Group, Inst. Of Biomedical Eng, Graz Univ. of Technology	(Java) http://genome.tugraz.at	Free	Low
GeneSense	InforSense	http://www.inforsense.com	Commercial	Low
GEPAS	Bioinf. Unit, National Spanish Cancer Center (CNIO)	(Web) http://gepas.bioinfo.cnio.es	Free	Medium
J-Express	MolMine	(Java) http://www.molmine.com	Academic license	Medium
MAExplorer	Open Source at SourceForge	(Java) http://maexplorer.sourceforge.net	Free/ Open source (MPL)	Medium
Microtracker	Ocimum Biosolutions	http://ocimumbio.com	Commercial	Medium
Partek Software Suites	Partek	(MS.W-Linux-Mac) http://www.partek.com	Commercial	Medium
PreP	University of Malaga, Spain	http://chirimoyo.ac.uma.es/bitlab/services/PreP/index.htm	Academic license	High
Soochika	Strand Genomics	(MS.W-Linux-Mac) http://www.strandgenomics.com	Upon request	Low
TIGR MIDAS	The Inst. Of Genom. Res. (TIGR)	(Java) http://www.tigr.org	Free/ Open source	Low
Xcluster	Standford Univ.	(MS.W-Linux-Mac) http://genome-www.stanford.edu/software	Free academic & non-profit	High
X-Miner	X-MINE	(Web) http://www.x-mine.com	Commercial	Low
Xpression	Informax Inc	(MS.W) http://www.informaxinc.com	Commercial	Low

Cluster & TreeView	Microarray Data Analysis Tools	http://rana.lbl.gov/index.htm	Free for academic	High
Qfirst	Quality is the first in gene expression data	www.integromics.com	Commercial	Low
GenMapp	Gene expression data visualization on maps representing biological pathways	http://www.genmapp.org/	OS	Low
ImaGene	Microarray/Macroarray Image Processing Tools	http://www.biodiscovery.com/	Commercial	Low
MeV	Microarray Data Analysis Tools	http://www.tm4.org/mev.html	Open Source	Low
MIDAS	Microarray Data Analysis Tools	http://www.tm4.org/midas.html	Open Source	Low
PathwayStudio	Literature/Database-based integration of microarray results with pathways.	http://www.ariadnegenomics.com/	Commercial	Low
Ingenuity Pathway Analysis (IPA)	Integration of microarray results with pathways	http://www.ingenuity.com/	Commercial	Low
Decision site	Spotfire	www.spotfire.com	Commercial	Low
TIGR SpotFinder	Microarray/Macroarray Image Processing Tools	http://www.tm4.org/spotfinder.html	Open Source	Low
SAGE	Serial Analysis of Gene Expression	http://www.sagenet.org/	OS	Low
Datamining tools				
MineBioText	MineBioText is a generic software tool for Data Mining in Biomedical Text Documents.	http://www.ics.forth.gr/~kantale/MineBioText/MineBioText.html Authors: antonak@ics.forth.gr, kantale@ics.forth.gr		Low
MineGene	MineGene is a general-purpose machine learning tool to serve as an application platform for various Data Mining Operations including gene selection, classification and clustering algorithms.	http://www.ics.forth.gr/~kantale/MineGene/MineGene.html Author: kantale@ics.forth.gr		Low

Association rule discovering	Software for disclosing associations among data.	A-priori based algorithms. (i.e "engene" at chirimoyo.ac.uma.es/engenet; Borgelt software at http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html)		Medium
ArrayUnlock	Association rule discovering between gene expression data and functional annotations	www.integromics.com	Commercial	Low
LitheMiner	Association rule discovering in scientific literature	www.integromics.com	Commercial	Medium
Ontology tools				
R - Bioconductor packages	R is a programming language and software environment for statistical computing and graphics. The bioinformatics community has seeded a successful effort to use R for the analysis of data from molecular biology laboratories. The bioconductor project started in the fall of 2001 provides R packages for the analysis of genomic data. e.g. Affymetrix and cDNA microarray object-oriented data handling and analysis tools.Availability of many tools for statistical analysis and very good for tool development and programming own routines – compatible with other programming languages (eg. C)	http://www.r-project.org/ http://www.bioconductor.org/	Open source	High
Data mediation / integration				
FACT++/Pellet	OWL-DL Reasoner	http://owl.man.ac.uk/factplusplus/ http://www.mindswap.org/2003/pellet/index.shtml	OS	Low
Jena Framework	Java framework for building Semantic Web applications. It provides a programmatic environment for RDF, RDFS and OWL, including a rule-based inference engine.	http://jena.sourceforge.net/	OS	Low
OWL	Ontology Web language	http://www.w3.org/TR/owl-features/		Low

Protégé 3.1	Ontology editor and knowledge-base framework.	http://protege.stanford.edu/ http://www.co-ode.org/resources/tutorials/ProtegeOWLTutorial.pdf	OS	Medium
FatiGO	Annotation by comparing lists of genes in terms of PubMed bio-entities (chimical and diseases), Gene Ontology terms (cellular component, biological process and molecular function), InterP,	http://bioinfo.cipf.es	Free	High
BioMart	Database federation search engine (mediation service)	http://www.biomart.org/	GPL	Low
PubMed	PubMed is a free search engine for the MEDLINE database. MEDLINE covers over 4,800 journals published in the United States and more than 70 other countries primarily from 1966 to the present.	http://pubmed.gov http://en.wikipedia.org/wiki/Pubmed		High
DAS systems	Distributted Annotation Servers (Reference and annotations servers). Several available Clients	http://www.ensembl.org/info/data/external_data/das/ensembl_das.html	free	High
Annotations (FunCUT)	Annotates homologous sequences including the identification of protein subfamilies (orthologous groups)	(http://www.pdg.cnb.uam.es/funcut.html) http://pdg.cnb.uam.es/das/funcut/ (DAS)		Low
bioBroker	XML-based mediator system (Swissprot, EMBL, KEGG and PDB)	uranos.khaos.uma.es/mediator	free	Medium
Proteomics				
Melanie & ImageMaster	2D gel image analysis software suite	Geneva Bioinformatics (GenBio): www.genebio.com	Commercial	Medium
Decyder (Amershan Pharmacia)	2D gel image analysis software suite	Amershan Pharmacia: http://www5.amershambiosciences.com	Commercial	Medium
Phoretix 2D (Phoretix International)	2D gel image analysis software suite	Phoretix International: http://www.nonlinear.com/	Commercial	Medium
Gellab	Image processing software and gel analysis systems	Scanalytics: http://scanalytics.com/product/gellab/2d-gel.shtml	Commercial	Medium
Kepler; Z3; GD-Impressionist	2D gel image analysis software suite	Large Scale Proteomics; Compugen; GeneData...	Commercial	Medium
Progenesis PG600	A new addition to the Progenesis range which has been developed for the processing and comparison of MS traces	http://www.nonlinear.com/products/progenesis/	Commercial	Medium

	from different sample classes and treatments allowing fast identification of potential protein biomarkers.			

27 Appendix 3- Important European Societies in Cancer Research

Name of Society or Organisation	
Cancer Research UK http://science.cancerresearchuk.org/ http://www.cancerhelp.org.uk/trials/trials/default.asp	Cancer Research UK is a cancer research and awareness-promotion group in the United Kingdom. It is the foremost cancer charity in the United Kingdom, and the biggest cancer research organisation outside the USA. It is accredited by the UK's National Health Service as a health information provider. Cancer Research UK supports and undertakes cancer research in hospitals, universities and medical schools throughout the United Kingdom, and disseminates information to the general public and the scientific community through its various websites, as well as its twice-monthly scientific publication, the British Journal of Cancer. It also makes information about current clinical trials accessible via its website; as at January 2006, there were 211 such trials open to UK cancer patients.
European Association for Cancer Research (EACR) http://www.eacr.org/about.html	The EACR was established in 1968 and the membership now exceeds 5000. The Association offers opportunities for: Communication o with laboratory, translational and clinical cancer researchers in all areas of oncology - from basic research, to prevention, treatment and care o at meetings and special conferences o with partner international cancer organizations o in scientific journal (European Journal of Cancer) o through the EACR Newsletters
European Association for Neurooncology (EANO) http://www.eano.de/	The EANO was finally established as an association in 1994. EANO is cooperating with national groups and stimulating international cooperation in neurooncologic research and clinical training. EANO is also involved in the responsibility for the World Conference of Neurooncology, which was first held in Washington in November 2002
European Organisation for Research and Treatment of Cancer (EORTC) http://www.eortc.be/	The EORTC is an international non-profit organisation that develops, coordinates and stimulates cancer laboratory and clinical research in Europe. It is located in Woluwe-Saint-Lambert/Sint-Lambrechts-Woluwe.

European Society of Gynaecological Oncology (ESGO) http://www.esgo.org/	The mission statement of the ESGO are as follows: ▪ To create an open European platform of individual professionals dedicated to the care of women with gynaecological cancer. ▪ To be the authority responsible for recognition of Gynaecological Oncology in Europe. ▪ To lead Europe in clinical and scientific education in Gynaecological Oncology and provide standards and supervision for certified training. ▪ To independently set multi-professional Standards of Care for women with gynaecological cancer. ▪ To integrate clinical and basic research into the educational, training and collaborative activities of the society. ▪ To promote communication with scientific and professional organisations within Europe and worldwide.
European Society for Medical Oncology (ESMO) http://www.esmo.org/	The ESMO is a professional organization representing medical oncologists. The Society focuses on a multidisciplinary approach to treatment and has expanded to include radiation and surgical oncologists, as well as other healthcare professionals involved in clinical cancer care. ESMO aims to unite physicians, caregivers, and patients in a global alliance committed to combating cancer and ensuring equal access to quality multidisciplinary treatment. ESMO strives to certify and maintain the highest standards of overall care for cancer patients by: ▪ gathering and disseminating oncology research results ▪ supporting oncologists and cancer patients with guidelines, policies and publications ▪ offering and accrediting state-of-the-art education and training programs and designated cancer centers
European Society of Surgical Oncology (ESSO) http://www.esso-surgeonline.be/	The ESSO was founded in 1981 to advance the art, science and practice of surgery for the treatment of cancer. By arranging scientific conferences, professional exchanges and seminars, ESSO endeavours to ensure that the highest possible standard of surgical treatment is available to cancer patients throughout Europe. It aims to foster multi-disciplinary collaboration in the clinical management of cancer patients. ESSO is increasingly involved in the training of surgeons concerned by cancer care throughout Europe and in promoting the development of guidelines of good practice in cancer surgery. The Society also seeks to promote knowledge and education about cancer care

	and to facilitate basic and clinical research in oncology.
European Society for Therapeutic Radiology (ESTRO) http://www.estro.be/	The ESTRO was founded in Milano in September 1980. as a Society of individual members working in the field of radiotherapy and oncology. Its principal objectives are to: ▪ Foster radiation oncology in all its aspects ▪ Develop standards for the quality assurance. of radiation oncology, radiophysics, radiation technology and radiobiology in Europe and stimulate their implementation ▪ Improve the standards of cancer treatment by establishing radiation oncology as a clinical specialty integrated with other cancer treatment modalities ▪ Promote international exchange of scientific information on radiotherapy & oncology and related fields of science such as radiophysics and radiobiology ▪ Set standards for education and practice in radiation oncology and associated professions ▪ Establish relationships and cooperation with international, regional and national societies and bodies in the field of radiation oncology.
Federation of European Cancer Societies (FECS) http://www.fecs.be/emc.asp	The FECS is the unique umbrella organisation gathering all the disciplines involved in research and treatment of cancer. This society is an international non-profit association that co-ordinates collaboration between European societies active in different fields of cancer research, prevention and treatment with the ultimate goal of providing the best possible treatment and care for all European cancer patients. Through its membership, FECS represents more than 18.000 experts involved in cancer research, treatment and care.
International Agency for Research on Cancer (IARC) http://www.iarc.fr/	The IARC is an intergovernmental agency forming part of the World Health Organisation of the United Nations. Its main offices are in Lyon, France. Its role is to conduct and coordinate research into the causes of cancer. It also conducts epidemiological studies into the occurrence of cancer worldwide.
International Society of Paediatric Oncology (SIOP) http://www.siop.nl/	The SIOP is the major global organisation concerned with the issues of children and young people who have cancer. For the past 35 years it has brought together doctors of many different disciplines to develop better care for this disease.
International Union Against Cancer (UICC)	As the world's largest independent, non-profit, non-governmental association of cancer-fighting organisations, UICC is a catalyst for responsible dialogue

http://www.uicc.org/	and collective action. UICC brings together a wide range of organisations, including voluntary cancer societies, research and treatment centres, public health authorities, patient support networks and advocacy groups. UICC's mission is to build and lead the global cancer control community engaged in sharing and exchanging cancer control knowledge and competence equitably, transferring scientific findings to clinical settings, systematically reducing and eventually eliminating disparities in prevention, early detection, treatment and care of cancers, and delivering the best possible care to all cancer patients.
Ludwig Institute for Cancer Research (LICR) http://www.licr.org/	The LICR is a global non-profit medical research institute that undertakes laboratory and clinical research into cancer, conducting and sponsoring its own early-phase clinical trials to investigate its discoveries. LICR is the largest international academic institute dedicated to understanding and controlling cancer, with ~900 staff in seven countries across Australia, Europe, and North and South America. There are currently nine LICR research Branches, which have a primary focus on basic laboratory and translational (in vivo and preclinical analyses of laboratory discoveries) sciences and are typically located within a university or research institute: o Brussels Branch of Human Cancer Cell Genetics o Lausanne Branch of Immunology o University College London Branch of Cell and Molecular Biology o Melbourne Branch of Tumour Biology o New York Branch of Human Cancer Immunology o San Diego Branch of Cancer Genetics o São Paulo Branch of Cancer Biology and Epidemiology o Stockholm Branch of Molecular and Cell Biology o Uppsala Branch of Growth Regulation

28 Appendix 4- National Cancer Research Centers

Belgium	Institut Jules Bordet, Centre des Tumeurs de l'Universite libre de Bruxelles	http://www.bordet.be
France	Institut National Du Cancer Instutut Gustave Roussy Instutut Curie	http://www.e-cancer.fr/ http://www.igr.fr http://www.curie.fr
Germany	Deutsches Krebsforschungsinstitut	http://www.dkfz.de/
Greece	'Metaxa' Cancer Hospital Of Piraeus Anticancer Oncological Hospital of Athens 'Saint Savvas'	http://www.metaxa-hospital.gr N/A
Italy	Centro di Riferimento Oncologico/Istituto Nazionale Tumori – IRCCS Istituto Nazionale per lo Studio e la Cura dei Tumori	http://www.cro.it http://www.istitutotumori.mi.it/
Japan	National Cancer Center	http://www.ncc.go.jp/
Netherlands	Nederlands Kanker Instituut - Antoni van Leeuwenhoek Ziekenhuis	http://www.nki.nl/
Poland	Polish Anti-Cancer Institute	N/A
Romania	N/A	N/A
Spain	Centro Nacional de Investigaciones Oncológicas (Spanish National Cancer Centre)	http://www.cnio.es/es/index.asp
Sweden	Karolinska Institutet	http://ki.se/
Switzerland	Swiss Institute for Experimental Cancer Research	http://www.isrec.ch/

United Kingdom	Cancer Research UK	http://science.cancerresearchuk.org/

A detailed list of members of the European Cancer Institutes can be found on the Webpage of the Organisation of European Cancer Institutes (OECI):

http://www.oeci-eeig.org/wcm453/index.php?option=com_content&task=section&id=8&Itemid=41.

29 Appendix 5: Members of ECL

BELGIUM Belgian Federation Against Cancer (www.cancer.be) 479 Chaussée de Louvain 1030 Brussels, Belgium Tel:+32 2 736 99 99 Fax:+32 2 734 92 50 E-mail: commu@cancer.be	Flemish League against Cancer (www.tegenkanker.net) Koningsstraat 217 B-1210 Bruxelles, Belgium Tel: + 32 (0) 2 227 69 69 Fax: + 32 (0) 2 223 22 00 E-mail: vl.liga@tegenkanker.be
CROATIA Croatian League Against Cancer Illica 197 10000 Zagreb, Croatia Tel:+385 1 3775 572 Fax:+385 1 3775 568	
CYPRUS The Cyprus Association of Cancer Patients and Friends (www.cancercare.org.cy) PO Box 23868 1687 Nicosia, Cyprus Tel:+357 2 345 444 Fax:+357 2 346 116 E-mail: caocpat1@cytanet.com.cy	The Cyprus Anti-Cancer Society (www.anticancersociety.org.cy) PO Box 25296 1308 Nicosia, Cyprus Tel: +357 2 249 7373 Fax: +357 2 231 6822 E-mail: vassilis.i@anticancersociety.org.cy
CZECH REPUBLIC League Against Cancer Prague (www.lpr.cz) Na Slupi 6 12842 Praha 2, Czech Republic Tel:+420 2 249 19 732 Fax:+420 2 2491 9732 E-mail: lpr@lpr.cz	
DENMARK Danish Cancer Society (www.cancer.dk) Strandboulevarden 49 DK-2100 ø Copenhagen, Denmark Tel:+45 35 25 75 00 Fax:+45 35 25 77 01 E-mail: info@cancer.dk	Faroese Association Against Cancer (www.krabbamein.fo) J. Paturssonargota 24 FO-100 Tórshavn, Faroe Islands Tel:+298 317 959 Fax:+298 315 727 E-mail: ffk@post.olivant.fo
FINLAND Cancer Society of Finland (www.cancer.fi) Liisankatu 21 B FIN-00170 Helsinki 17, Finland Tel:+358 9 135 331 Fax:+358 9 135 1093 E-mail: society@cancer.fi	**FRANCE** Ligue Nationale contre le Cancer (www.ligue-cancer.asso.fr) 14 rue Corvisart 75013 Paris, France Tel: +33 1 53 55 24 00 Fax: +33 1 43 36 91 10 E-mail: naudc@ligue-cancer.net
GERMANY Deutsche Krebshilfe (www.krebshilfe.de)	Deutsche Krebsgesellschaft (www.krebsgesellschaft.de)

Postfach 1467 53004 Bonn, Germany Tel: +49 228 72990-0 Fax: +49 228 72990-11 E-mail: deutsche@krebshilfe.de	Steinlestrasse 6 D-60596 Frankfurt, Germany Tel:+49 69 63 009 60 Fax:+49 69 63 00 96 66 E-mail: beck@krebsgesellschaft.de
GREECE Hellenic Cancer Society (www.cancer-society.gr) A. Tsoha 18-20 GR-11521 Athens, Greece Tel:+30 1 6456 713 – 715 Fax:+30 1 6410 011 E-mail: hellas-cancer@ath.forthnet.gr	Hellenic Society of Oncology 11 Valtetsiou St. 10680 Athens, Greece Tel:+30 1 362 5637 Fax:+30 1 3611 774 E-mail: HSO@ath.forthnet.gr
HUNGARY Hungarian League Against Cancer (www.rakliga.hu) Post Box 7 1507 Budapest, Hungary Tel:+36 1 225 16 21/22,23 Fax:+36 1 202 4017 E-mail: rakliga@matavnet.hu	**ICELAND** The Icelandic Cancer Society (www.krabb.is) P.O.Box 5420 IS-125 Reykjavik, Iceland Tel:+354 540 1900 Fax:+354 540 1910 E-mail: gudrunag@krabb.is
IRELAND Irish Cancer Society (www.cancer.ie) 5 Northumberland Road Dublin 4, Ireland Tel: +353 1 2310 500/Fax: +353 1 2310 555 E-mail: reception@irishcancer.ie	
ITALY AIMaC Associazione Italiana Malati di Cancro, parenti i amici (www.aimac.it) Via Barberini 11 I-00187 Rome, Italy Tel: +39 06 482 5107 Fax: +39 06 4871 492 E-mail: info@aimac.it	Lega Italiana per la lotta Contro I Tumori (www.legatumori.it) Via A. Torlonia, 15 I-00161 Rome, Italy Tel:+39 06 44 25 971 Fax:+39 06 44 25 97 32 E-mail: sede.centrale@legatumori.it
LUXEMBOURG Fondation Luxembourgeoise (www.cancer.lu) contre le Cancer 209, route d´Arlon L-1150 Luxembourg Tel:+352 45 30 33 1 Fax:+352 45 30 3333 E-mail: flcc@pt.lu	
THE NETHERLANDS Association of Comprehensive Cancer Centres (www.ikc.nl) P.O Box 330 9700 AH Groningen, The Netherlands Tel. +31 50 521 59 00 Fax +31 50 521 59 99 E-mail: r.otter@ikn.nl	Dutch Cancer Society (www.kankerbestrijding.nl) Sophialaan 8 1075 BR Amsterdam, The Netherlands Tel:+31 20 57 00 550 Fax:+31 20 67 50 302 E-mail: info@kankerbestrijding.nl

POLAND Polish Anti-Cancer Committee 5 Roentgena Street 02-781 Warsaw, Poland Tel:+48 22 643 93 79 Fax: +48 22 643 90 63 E-mail: zwronkowski@coi.waw.pl	**SLOVAKIA** League Against Cancer in Slovakia (www.lpr.sk) Spitalska 21 812 32 Bratislava, Slovakia Tel: Fax: +42 1 252 921 735 E-mail: lpr@rainside.sk
PORTUGAL Liga Portuguesa contra o Cancro (www.ligacontracancro.pt) Av. Columbano Bordalo Pinheiro, 57-30 F 1070-061 - Lisboa Portugal Tel: +351 21 722 1810/ Fax: +351 21 726 8059 E-mail: info@ligacontracancro.pt	
SLOVENIA The Association of Slovenian Cancer Societies Zaloska 2 1000 Ljubljana, Slovenia Tel:+386 01 4309 780 Fax:+386 01 4309 785 E-mail: MZakelj@onko-i.si	**SWITZERLAND** Swiss Cancer League (www.swisscancer.ch) Effingerstrasse 40 CH-3001 Bern, Switzerland Tel:+41 31 389 91 14 Fax:+41 31 389 91 60 E-mail: info@swisscancer.ch
TURKEY Turkish Association for Cancer Research and Control Atac Sokak No:21 06420 Yenisehir-Ankara Tel:+90 312 431 29 50 Fax:+90 312 431 39 58 E-mail: tkutluk@tr.net	
UNITED KINGDOM Cancer Research UK (www.cancerresearchuk.org) 61 Lincoln´s Inn Fields London WC2A 3PX United Kingdom Tel:+44 20 7242 0200 Fax:+44 20 7269 3100 E-mail: claire.mallinson@cancer.org.uk	Macmillan Cancer Relief (www.macmillan.org.uk) 89 Albert Embankment SE1 8UQ London United Kingdom Tel: +44 020 7840 7840 Fax: +44 020 7840 7841 E-mail: jsimpson@macmillan.org.uk
Ulster Cancer Foundation (www.ulstercancer.org) 40-42 Eglantine Avenue BT9 6DX Belfast, Northern Ireland Tel:+44 2890 66 3281 Fax:+44 2890 66 0081 E-mail: ucfinfo@ulstercancer.org	

30 Appendix 6 - Abbreviations and acronyms

Bio-medical glossary

Adjuvant

After surgery.

Angiogenesis

The physiological process involving the growth of new blood vessels from pre-existing vessels. (Source [1])

Antigen

A substance that stimulates an immune response, especially the production of antibodies. (Source [2])

Apoptosis

One of the main types of programmed cell death (PCD). It is carried out in an ordered process that generally confers advantages during an organism's life cycle. (Source [3])

ADR

Adverse Drug Reaction.

Asymptomatic

Without symptoms. Showing no signs (of a disease).

Biopsy

Removal of a sample tissue from a living body for diagnostic purposes.

Carcinoma

A malignant tumor that begins in the lining layer (epithelial cells) of organs. At least 80% of all cancers are carcinomas. (Source [4])

cDNA library

Represents a complete, or nearly complete, set of all the mRNAs contained within a cell or organism. (Source [5])

Cell line

A population of cells propagated in culture that are totally derived from, and therefore genetically identical to, a single common ancestor cell. (Source [6])

Concomitant

Occurring simultaneously. Refers to symptoms that happen at the same time.

CRF

Case Report Form. Form used for recording data during a clinical trial.

Cystic

Closed sac containing air, fluids, or semi-solids.

DNA

Deoxyribonucleic acid. The molecule that encodes genetic information.

DNA microarray

A collection of microscopic DNA spots attached to a solid surface, such as glass, plastic or silicon chip forming an array for the purpose of expression profiling. (Source [7])

Efficacy

The ability to produce a desired amount of a desired effect. In a medical context it indicates that the therapeutic effect of a given intervention is acceptable. (Source [8])

Endocrine system

A control system of ductless glands that secretes hormones which circulate within the body via the bloodstream to affect distant cells within specific organs. (Source [9])

Epigenetic

The study of reversible heritable changes in gene function that occur without a change in the sequence of nuclear DNA. (Source [10])

Etiology

The study of causation. In medicine in particular, the term refers to the causes of diseases or pathologies. (Source [11])

Exon

A region of DNA within a gene that are is spliced out from the transcribed RNA and is retained in the final messenger RNA (mRNA) molecule. (Source [12])

Expression profiling

A technique of measuring the expression of particular genes, typically performed using microarray technology. See [13]

Gene expression

The process by which a gene's DNA sequence is converted into the structures and functions of a cell. (Source [14])

Gene regulation

Regulation of gene expression. The cellular control of the amount, timing and appearance of the functional product of a gene. (Source [15])

Germline

The line (sequence) of germ cells of a mature or developing individual that have genetic material that may be passed to a child. (Source [16])

Hemorrhage

The medical term for bleeding.

Heterozygosity

Carrying different versions of a gene. One can be recessive and is therefore not expressed in the heterozygous state.

Histology

The study of tissue sectioned as a thin slice. (Source [17])

Histopathology

The field of pathology which specialises in the histologic study of diseased tissue. (Source [18])

Humoral immunity

The aspect of immunity that is mediated by secreted antibodies, produced in the cells of the B lymphocyte lineage (B cell). (Source [19])

Immunogenic

Provoking an immune response when introduced into the body. (Source [20])

Immunoscreening

A method of biotechnology to detect a polypeptide produced from a cloned gene. (Source [21])

IMP

Investigational Medicinal Product

In vitro

The technique of performing a given experiment in a test tube or, generally, in a controlled environment outside a living organism. (Source [22])

In vivo

That which takes place inside an organism. (Source [23])

Intron

Sections of DNA that will be spliced out after transcription, but before the RNA is used. C.f. exon. (Source [24])

Lesion

A localised area of diseased or disordered tissue. (Source [25])

Library

A library is a collection of molecules in a stable form that represents some aspect of an organism. (Source [26])

LOH

Loss of Heterozygosity. In a cell, it represents the loss of a single parent's contribution to part of the cell's genome. A common occurrence in cancer, it often indicates the presence of tumor suppressor gene in the lost region. (Source [27])

Mastectomy

The surgical removal of one or both breasts, partially or completely. (Source [28])

Metabolic pathway

A metabolic pathway is a series of chemical reactions occurring within a cell, catalyzed by enzymes, resulting in either the formation of a metabolic product to be used or stored by the cell, or the initiation of another metabolic pathway. (Source [29])

Metabolism

The biochemical modification of chemical compounds in living organisms and cells. It includes the biosynthesis of complex organic molecules (anabolism) and their breakdown (catabolism). (From [30])

Metabolite

An intermediate or end-product of metabolism. The term metabolite is usually restricted to small molecules. (Source [31])

Metastasis

The spread of cancer from its primary site to other places in the body (e.g., brain, liver). (Source [32])

Microarray

Refers to DNA microarrays, protein microarrays, antibody microarrays or other biological assays depending on the context. (Source [33])

Necrosis

Unprogrammed or accidental death of cells and living tissue. (Source [34])

Necrosis of the tumour

No vital tumour cells remaining. Surgically removed tumour remains only consist of fibre and other non-cell material.

Neoadjuvant

Pre-operative (before surgery)

Neoplasm

An abnormal, disorganized growth in a tissue or organ, usually forming a distinct mass. Neoplasms may be benign or malignant. Also called tumor. (Source [35])

Nephrectomy

The surgical removal of a kidney. (Source [36])

Palable

Perceptible by touch

Pathogenesis

The mechanism by which a certain etiological factor causes disease. (Source [37])

Pathognomonic

Characteristic or diagnostic for a particular disease. (Source [38])

Peptide

The family of short molecules formed from the linking, in a defined order, of various α-amino acids. (Source [39])

Pharmacogenomics

The branch of pharmaceutics that deals with the influence of genetic variation on drug response in patients by correlating gene expression or single-nucleotide polymorphisms with a drug's efficacy or toxicity. (Source [40])

Presymptomatic

Having a disease, but without the symptoms yet.

Promoter

A DNA sequence that enables a gene to be transcribed. (Source [41])

Proteome

The entire complement of proteins in a given biological organism or system at a given time, i.e. the protein products of the genome. (Source [42])

Proteomics

The large-scale study of proteins, particularly their structures and functions. It includes the way they work and interact with each other inside cells. (See [43])

Post genomic

Teasing higher biological meaning and function out of raw sequence data. (Source [44])

RNA

Ribonucleic Acid. A molecule found in the nucleus and cytoplasm of cells. It plays an important role in protein synthesis. (Source [45])

Renal tumour

Tumour in the kidney

SAE

Severe Adverse Effects

SEREX

Serological expression of cDNA expression libraries. A technique for identifying antibody products in serum. See [46] for more details.

Serology

The branch of science dealing with properties and reactions of sera, particularly the use of antibodies in the sera to examine the properties of antigens.

Serum

The clear liquid part of the blood that remains after blood cells and clotting proteins have been removed. (Source [47])

SNP

Single Nucleotide Polymorphism. A DNA sequence variation that occurs when a single nucleotide (A,T,C,or G) in the genome sequence is altered. (Source [48])

Stoichiometry

The calculation of quantitative (measurable) relationships of the reactants and products in chemical reactions (chemical equations). (Source [49])

Stratification

Separation of a study cohort into subgroups or strata according to specific characteristics, so that these differences which might affect the outcome of the study, can be taken into account. (Source [50])

SUSAR

Suspected Unexpected Severe Adverse Reactions

Systemic

Spread throughout the body; affecting many or all body systems or organs; not localized in one spot or area. (Source [51])

Transcription

The process through which a DNA sequence is enzymatically copied by an RNA polymerase to produce a complementary RNA. (Source [52])

Tumor-associated antigens

Antigens that are presented by tumor cells and normal cells. C.f. tumor-specific antigens (Source [53])

Tumor-specific antigents

Antigens that are specific to tumor cells. (Source [54])

Bio-medical technologies

BLAST

Basic Local Alignment Search Tool. This tool finds regions of local similarity between sequences. See [1].

BioMOBY

A biological web service interoperability initiative. See [2]

caGrid

Grid software for making cancer related bio-medical data accessible across institutions. Part of caBIG. See [3].

CDA

Clinical Document Architecture. An HL7 standard for marking up documents for exchange.

CDISC

Clinical Data Interchange Standards Consortium. See [4].

CGL7

Clinical Genomics Level Seven.

DICOM

Digital Imaging and Communications in Medicine. See [5]

EMBOSS

An Open Source software analysis package for molecular biology. See [6]

GALEN

A technology that is designed to represent clinical information in a new way, using a qualitatively different approach from those used in the past. See [7].

Gene Ontology (GO)

an ontology that describes biological processes, cellular components, and molecular functions. See [8].

HL7

Health Level Seven. See [9]

MAGE

Standard for exchanging micro-array expression data. See [10].

MAML

Microarray Markup Language. Superseded by MAGE-ML.

MIAME

Minimum Information About a Microarray Experiment. See [11].

RIM

Reference Information Model. An HL7 standard.

SNOMED

Systematized Nomenclature of Human and Veterinary Medicine.

Swiss-Prot

A curated protein sequence database.

Taverna

Workflow tool for instantiating bio-informatics webservices. Produced by the myGrid project.

TrEMBL

A computer-annotated supplement of Swiss-Prot.

UDIP

Universal De-Identification Platform by IBM Research. See [12].

UMLS

Unified Medical Language System. See [13].

UniProt

The most comprehensive catalog of information on proteins. See [14].

UniProtKB

UniProt Knowledge Base.

WADO

Web Access to DICOM Objects. A standard that is part of the DICOM specification.

Technical Glossary

Condor

Job management (distribution and monitoring) for Grids.

DAML+OIL

DARPA Agent Markup Language - Ontology Integration Language. Web Ontology Language (predecessor of OWL?).

DFDL

Data Format Description Language. Defined as part of OGSA–DAI.

GAT

Grid Application Toolkit. Produced as part of the GridLab project. See [1]

GGF

Global Grid forum. See [2]

Globus toolkit

An open source software toolkit used for building Grids. See [3]

GridFTP

Provides file transfer within Grid. Defined by GGF.

Gridge

Grid middleware developed by PSNC. See [4].

GridLab

A Grid application Toolkit and Testbed. IST project with PSNC as a participant. See [5].

GridSuite

Grid middleware developed by PSNC. Now called Gridge.

GSI

Grid Security Infrastructure. Part of Globus.

GRAIL

Description logic dialect. Used to specify GALEN ontology (not widely used elsewhere?).

GRAM

Globus Resource Allocation Manager. A Globus service for submitting and controlling jobs on heterogeneous Grid computation elements.

GRMS

GridLab Resource Management System. Developed as part of the GridLab project. See [6].

ISO/IEC 11179

Information Technology -- Metadata Registries (MDR). See [7]

MDR

Metadata Registries. An ISO standard: ISO/IEC 11179.

MPI

Message Passing Interface. Low-level interfaces for efficient distributed messaging.

MyProxy

A Globus network service that stores user credentials so they can be accessed from other systems on the network.

OASIS

Organization for the Advancement of Structured Information Standards.

OGSA

Open Grid Services Architecture. Defined by GGF.

OGSA–DAI

Data Access and Integration. A middleware product that supports the exposure of data resources, such as relational or XML databases, on to Grids. See [8].

OGSI

Open Grid Services Infrastructure. Defined by GGF.

OntoGrid

An IST project that aims to provide technological infrastructure for the rapid prototyping and development of knowledge-intensive distributed open services for the Semantic Grid. See [9].

OWL

Web Ontology Language. See [10], [11].

RDF

Resource Description Framework. Intended for representing metadata about Web resources.

RFT

Reliable File Transfer. An RFT service is provided by Globus.

RLS

Replica Location Service. Part of Globus.

RTF

Reliable File Transfer (bloody French ;-). A protocol for reliably transfering files.

Shibboleth

A set of Globus services that leverage existing user authentication and authorization systems at "home institutions" to give remote services the information they need to make authorization decisions.

SOAP

XML-based messaging for web services

UDDI

Web services directory

WSDL

XML-based service description. It includes a definition of data types, messages, port-types (operations), binding (transmission details), service (location).

WSRF

WS-Resource Framework. See [12], [13].

Legal and ethical glossary

Admissibility of data processing

The collection, processing and use of personal data shall be admissible only

- if permitted or prescribed by law or
- if the data subject has consented.

This is the basic message of data protection law. If ACGT processes personal data one of these two exceptions must be corresponding. Therefore it should be an aim for ACGT to process as little personal data as possible.

Aliasing (Pseudonymising)

Aliasing means replacing a person's name and other identifying characteristics with a label, in order to preclude identification of the data subject or to render such identification substantially difficult. We assume that most data processed in ACGT is aliased. Aliased data still is “personal data” in the legal sense. As the differentiation between anonymous data and aliased data is both critical and difficult, please contact us to check what kind of data you are working with.

Anonymous data / Rendering anonymous

Rendering anonymous means the modification of personal data so that the information concerning personal or material circumstances can no longer or only with a disproportionate amount of time, expense and labour be attributed to an identified or identifiable individual. Personal data that was anonymized is no longer “personal data” in the legal sense. It will have to be an aim to have as much anonymized data within ACGT as possible and reasonable.

Automated decision

Every person has the right according to Art. 15 (1) Directive 95/46/EC not to be subject to a decision which produces legal effects concerning him or significantly affects him and which is based solely on automated processing of data intended to evaluate certain personal aspects relating to him, such as his performance at work, creditworthiness, reliability, conduct, etc (automated decision). Automated decisions are allowed, if that decision is taken in the course of the entering into or performance of a contract, provided the request for the entering into or the performance of the contract, lodged by the data subject, has been satisfied or that there are suitable measures to safeguard his legitimate interests, such as arrangements allowing him to put his point of view. Such automated decisions are also permitted, if they are authorized by a law which also lays down measures to safeguard the data subject's legitimate interests. Every data subject has then the right to know the logic involved in

the automatic processing of data concerning him (Art. 12 (a) Directive 95/46/EC). For ACGT this means that all decisions that produce legal effects on a person should generally be made by an individual person and not by a computer or any other data processing system.

Coded (encrypted) data

Coded data is encrypted data. If it is personal data it can only be linked directly or indirectly to a natural person through a code. In ACGT, we guess the data will be coded. The data concerning a data subject shall be either encrypted by a code and/or via an alias.

Confidentiality

Persons employed in data processing shall not collect, process or use personal data without authorisation (confidentiality). On taking up their duties such persons shall be required to give an undertaking to maintain such confidentiality. This undertaking shall continue to be valid after termination of their activity. Any person acting under the authority of the controller or of the processor, including the processor himself, who has access to personal data must not process them except on instructions from the controller, unless he is required to do so by law. Researcher in the context of ACGT are therefore only allowed to collect, process and use personal data of a patient in compliance with the patient's informed consent. They are not allowed to disclose any data, unless they are authorised by the particular patient.

Consent

The data subject's consent means any express indication of his wishes by which the data subject signifies his agreement to personal data relating to him being processed, on condition he has available information about the purposes of the processing, the data or categories of data concerned, the recipient of the personal data, and the name and address of the controller and of his representative if any. The data subject's consent must be freely given and specific, and may be withdrawn by the data subject at any time. Concerning ACGT we assume that the specification of the consent will be critical and might create large databases that have to be managed. If the data subject is incapable of a free decision or domestic laws don't permit the data subject to act on his/her own behalf, consent is required of the person recognised as legally entitled to act in the interest of the data subject or of an authority or any person or body provided for by law. An informed consent of the particular patient is a vital requirement in order to collect and use the data needed for ACGT lawfully, though it is not the only possibility. The processing of personal data can be permitted expressively by law also. If the data subject is a minor (which will be the regular case in the Nephroblastoma-studies), the informed consent of the legally entitled persons, normally the minor's parents, is needed.

Data controller

The controller is, according to the Data Protection Directive 95/46 CE, the natural or legal person who alone, or jointly with others, determines the purposes and means of the processing of personal data. It is important to identify who the controller of any processing is, since this controller is the one liable for the legality of the processing and the fulfilment of the obligations towards the national data protection authority and the data subjects. The determination of the controller is factual and, in ACGT, will depend on the analyse of the system and the data flows.

Data processor

Data processor shall mean a natural or legal person, public authority, agency or any other body which processes personal data on behalf of the controller who is liable for the legality of the processing and the fulfilment of the obligations towards the national data protection authority and the data subjects. In ACGT it will be important to know, which partner within the project acts only as a data processor and who acts as a data controller as most of the data protection duties are relevant for the data controller and not the data processor.

Data reduction/Data economy (Minimality)

Personal data must not be excessive in relation to the purposes for which they are collected and/or further processed. It is therefore not allowed to process any data unless the data is necessary to achieve the purpose mentioned for which the data are collected and further processed. In case the processing of data is needed, only as little personal data as possible should be processed. The processed personal data has to be erased or anonymised once they are no longer required for the purposes for which they have been kept. For ACGT this means that it is only allowed to process (collect, use etc.) this kind of personal data of a patient that is needed for this project.

Data Subject

The data subject is the subject of personal data, i.e. an identified or identifiable person about whom the personal data refers. An identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity. Regularly the patient, whose genomic data is collected and used for the ACGT-studies, will be the data subject.

Disclosing

The disclosure of personal data to third parties or recipients is a processing operation and, as such, is subject to the legal requirements of processing. The rule for the technical and organizational requirements is the confidentiality of the personal data. Therefore, the controller must ensure the confidentiality of personal data, meaning that unauthorized access to, or disclosure of, the personal data, must be prevented. If there is a disclosure to a third party or a recipient, the controller should check whether or not this transfer or disclosure falls within the scope of the initial purpose or is still compatible with this purpose, in order to determine whether or not they can transfer or disclose the data. Anonymous data can be transferred without being subject to specific requirements. It's, also, used to fix some delay for the execution of obligation. For example, the controller (or his representative) must provide the required information to the data subject, if disclosure to a third party is anticipated, no later than the time when the data are first disclosed, except when the data subject has already been provided with the information.

Modification

The modification of personal data is considering by the Data Protection Directive 95/46 CE as part of the processing and concerns different things as the rectification, erasure and blocking. The data subject has the right to obtain from the controller the rectification, erasure or blocking the data processing because of the incomplete, inaccurate nature or illegal processing of the data.

Necessary processing

When deciding which data will be collected and further processed, the controller must limit these data to the extent strictly necessary to achieve the purpose of processing. This means that personal data will only be processed when it is necessary for the project.

Obtaining/Collecting

Collecting or obtaining the data is considered by the Data Protection Directive 95/46 CE as part of the processing. We use the term of:

- primary collection when the collection of personal data is directly obtained from the data

subject, ie either directly provided by the data subject or obtained through observation of the data subject.

- secondary collection when the collection of personal data is obtained from sources other

than the data subject himself. ACGT will deal with both primary and secondary collections because a collection will often be re-used for another purpose than the first one.

Organizational measures

Organizational measures must ensure combined with technical measures an appropriate level of security of the data processing, taking into account the state of the art and the costs of their implementation in relation to the risks inherent in the processing and the nature of the data to be protected. Appropriate organisational measures shall be taken by the controller against accidental loss, destruction or alteration of, or damage to, personal data and against unauthorised or unlawful processing of personal data in particular where the processing involves the transmission of data over a network, and against all other unlawful forms of processing. The controller must, where processing is carried out on his behalf, choose a processor providing sufficient guarantees in respect of the technical security measures and organizational measures governing the processing to be carried out, and must ensure compliance with those measures. Such appropriate organizational measures to ensure the confidentiality, integrity and accuracy of processed data should be for example:

- control of the entrance to installations
- control of data media
- memory control
- control of utilisation
- access control
- control of communication
- control of data introduction

- control of transport
- availability control

Such organizational measures have to be taken by all the ACGT-participants processing personal data. It should be determined prior to processing of personal data, which person is responsible for each organizational measure and by which means these goals can be achieved.

Personal data

Personal data means any information relating to an identified or identifiable natural person ('data subject'). An identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity. Therefore a set of date collected under a certain number or sign “patient xxx”, “tissue YYY” can be personal data.

Processing/Automated Processing

The concept of processing is very broad. It concerns any operation or set of operations that are performed upon personal data, whether or not by automatic means. Data processing is considered to be the collection, recording, organisation, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available (eg by allowing the inspection of data retrieval by a third party), alignment or combination, blocking, erasure or destruction of personal data. The application of data protection legislation is limited to automated processing and to nonautomated processing. Both types of processing operations form part of a filing system or are intended to form part of a filing system, ie any structured set of personal data which are accessible according to specific criteria, whether centralised, decentralised or dispersed on a functional or geographical basis. The processing operations covered by data protection legislation are therefore not limited to electronic files or databases, but also include the processing of data in a manual paper file as soon as this is structured according to certain criteria. The concept of processing1 also includes the operations performed by Internet software and hardware without the knowledge of the data subject, and hence invisible to them, such as the use of cookies. The exchange of information related to the use of browser software is also to be considered as processing. Obviously, ACGT processes data. 1 According to Group 29.

Public Register

A public register is a register which according to laws or regulations is intended to provide information to the public and which is open to consultation either by the public in general or by any person who can demonstrate legitimate interest, to the extent that the conditions laid down in law for consultation are fulfilled in the particular case.

Publish

The controller should refrain from publishing personal data or otherwise making them public. In most cases this will not be necessary to achieve the purpose of the research, or it may create an attempt to the data subject’s interests that appears to be disproportionate to the interest of the controller. The notion of making public is also a criteria to allow the processing. It's the case when the data subject has manifestly

made public his personal date concerning, for example, his health, the processing is allowed (article 8.2.e of the Directive 95/46/EC).

Purpose

The purposes for processing of personal data must be adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed. The purposes must be specified, explicit and legitimate. Personal data must be not further processed in a way incompatible with those purposes. It will be a critical task for ACGT to define the limits of purpose given by existing data and existing data processings.

Sensitive (personal data)/Special categories of data

Sensitive personal data is personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and data concerning health (genomic data) or sex life. Member States shall prohibit the processing of these data, except in explicitly stated exceptions. A lot of personal information in ACGT will be sensitive data which means that any data processing needs to be covered by an exceptional permissibility.

Statistical processing

Statistical processing is any operation of collection and processing of personal data necessary for statistical surveys or for the production of statistical results. These statistical results may further be used for different purposes, including a scientific purpose. The statistical purpose cannot lead to the possibility of taking individual decisions.

Storage

Storage of personal data is allowed by the Data Protection Directive 95/46 CE. BUT when the purpose of processing is achieved, and the data are not required any more for that particular purpose, these personal data must be rendered anonymous or be destroyed. Most national laws allow personal data to be stored for a longer term, provided that this is in order to use the data exclusively to carry out scientific research or statistics. Nevertheless, some national laws impose supplementary conditions or formalities in order to allow longer storage.

Third Party

The third party is a natural or legal person, public authority, agency or any other body other than the data subject, the controller, the processor and the persons who, under the direct authority of the controller or the processor, are authorised to process the data. In ACGT, the third party will be the other centres of researches, administration, other hospitals, etc…

Transfer (also to Third Countries)

The purpose of the Data Protection Directive 95/46 CE is to allow the free flow of personal data between Member States. The other objective of the Directive is to protect the fundamental rights and freedoms of natural persons, and in particular their right to privacy with respect to the processing of personal data. The Directive defines specific conditions and restrictions guaranteeing the protection of data subjects, but

the Member States are not allowed to restrict or prohibit these flows to a greater extent than permitted in the framework of the Directive. A specific regime regarding the transfer of personal data to non-EEA countries has been put in place to protect the data subjects whose data are exported outside the territorial scope of the application of the Directive. Before transferring data to a third country, the controller must check if the third country allows an adequate level of protection. If it's not so, the transfer can't take place except some exceptions mentioned in the article 25 of the Directive 95/46 CE: (a) the data subject has given his consent unambiguously to the proposed transfer; or (b) the transfer is necessary for the performance of a contract between the data subject and the controller or the implementation of precontractual measures taken in response to the data subject's request; or (c) the transfer is necessary for the conclusion or performance of a contract concluded in the interest of the data subject between the controller and a third party; or (d) the transfer is necessary or legally required on important public interest grounds, or for the establishment, exercise or defence of legal claims; or etc… The word "transfer" concerns also the disclosure of the data to a recipient or third person (see those words). Using The use of personal data is considering by the Directive 95/46 CE as part of the processing and is a criteria used to define the processing and the scope.

Recipient

The recipient is a natural or legal person, public authority, agency or any other body to whom data are disclosed, whether a third party or not. Authorities that may receive data during a particular inquiry shall not be regarded as recipients (article 2 of the Directive 95/46/EC)

Recording

Recording is a process and a criteria to determine the scope of the Directive 95/46/EC. The Directive uses it to fix some delay for the execution of obligations. For example, the controller (or their representative) must provide the required information to the data subject at the latest at the time of recording, except when the data subject has already been provided with the information (article 11).