Consolidated Requirement Analysis for Data Mining, Analysis and the Visualization Environment

Project Number: FP6-2005-IST-026996

Deliverable id: D6.1

Deliverable name: Consolidated Requirement Analysis for Data Mining, Analysis and the Visualization Environment

Date: 24 January 2007

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 6.1
Document name:	Consolidated Requirement Analysis for Data Mining, Analysis and the Visualization Environment
Document type (PU, INT, RE)	INT
Version:	2.0
Date:	24.01.2007
Authors: Organisation: Address:	WP6 Partners

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

ABSTRACT:

The present document provides an overview of the contributions of individual partners in ACGT WP6. The contributions are as follows: SIB provides datasets based on actual clinical-trial for testing, FhG addresses data mining tools and use of R in a grid environment, LundU addresses database developments and analysis tools based on pathway information, Biovista proposes a literature-based analysis tool, INRIA develops new clustering algorithms, UPM addresses the use of ontologies in the clinical data analysis, and UvA proposes a framework for interactive visualization.

KEYWORD LIST: Knowledge and data management, discovery tools, data mining.

MODIFICATION CONTROL			
Version	Date	Status	Author
0.1	07.07.2006	Draft	WP6 Partners (ed. Thierry Sengstag)
0.2	12.07.2006	Draft	WP6 Partners (ed. Thierry Sengstag)
0.4	15.10.2006	Draft	Stefan Rüping
0.5	27.10.2006	Draft	Thierry Sengstag (shuffling paragraphs)
1.0	24.11.2006	Draft	Stefan Rüping
2.0	24.01.2007	Final	Stefan Rüping

List of Contributors

- Stefan Rüping, FhG (editor)
- Alberto Anguita, UPM
- Tallur Basavanepa, INRIA
- Robert Belleman, UvA
- Francesca Buffa, UOXF
- Federico García, UMA
- Jari Häkkinen, LundU
- Israël Cesar Lerman, INRIA
- Andreas Persidis, Biovista
- Thierry Sengstag, SIB
- Alejandro Sola, UMA
- Oswaldo Trelles, UMA

Contents

Executive Summary

The goal of this deliverable is to consolidate the user requirements defined in Deliverable 2.1, concerning the knowledge mining and the visualization environment of the ACGT platform. It will complement the top-down approach that was used in the elicitation of the user requirements by a bottom-up approach, with the goal of pushing available analysis technologies to the use in the mining of clinico-genomic data. In addition, requirements that are implicitly posed to the analysis environment by key architectural decisions, such as the Grid approach, or general requirements like easy usability by non-experts, will be documented.

In particular, requirement consolidation in this document concerns the requirements for specific biomedical analysis tools, the requirements for generic, grid-enabled knowledge discovery tools, knowledge management and visualization. In addition, the agreed-upon approach for the elicitation of additional requirements, their implementation and testing are described.

1 Introduction

3.1 Introduction

The aim of the ACGT infrastructure is to facilitate clinical research and trials. In this context WP6 will provide the tools and infrastructure related to the scientific exploitation of clinico-genomics data. The framework for the development and implementation of WP6 tools is clearly defined by the needs of three types of user groups, namely end-users (primarily clinicians), biomedical researchers and data miners. Each user group is defined by a different focus (e.g. patients vs. trials vs. sets of trials), different levels of bio-medical knowledge and data analysis expertise, and different research interests (e.g. understanding a certain type of cancer vs. understanding the merits of a certain data analysis approach).

Starting from a very general definition of software quality, the user requirements for the ACGT data analysis environment can be divided into the following aspects:

- **Appropriateness**: the data analysis environment should provide the appropriate tools and services to support users in the state-of-the-art scientific analysis of biomedical data. Section 4 gives a list of specific tools and services that will be supplied by the platform.
- **Extensibility and reusability**: the platform should be easily extensible to new tasks and existing solutions should be easily reusable and transferable to similar analysis problems. This goal will be approached by supporting a semantically rich description of services and workflows, including a repository for the storage and retrieval of workflows (see Section 3), plus a tool for workflow recommendation based on a semantic description of the data (Section 4).
- **Performance**: the system must be performant enough to facilitate large analysis and optimization tasks, which calls for an efficient use of the Grid architecture. Section 2 will give details about the requirements for Grid-based data mining.
- **Usability**: the system should be easy to use for inexperienced users, but also provide a powerful interface for experts. Section 1.5 presents the requirements for the user interfaces for each of the different user types.
- **Security**: The system must be secure and protect the privacy of the involved patients. However, privacy will be dealt with in the context of another work package, and hence this aspect will not be discussed in this deliverable.

From the viewpoint of the end-user, the user requirements are better expressed as a set of typical clinico-genomics scenarios which can be used as a guide for the development process. For instance, the scenarios listed in D2.1 are:

SC1: A Compex Query Scenario for the TOP Trial

This scenario, which has a strong emphasis on high-throughput genomics (microarrays), covers a large number of typical actions performed in the context of the research activities associated to clinical trial. This covers: data storage and retrieval, preprocessing, data analysis, genomic annotation, pathway analysis, literature search and mining, and reporting.

SC2: Identification of Nephroblastoma Antigens

This scenario aims at the identification of immunogenic tumor-associated antigens in risk-stratified groups of patients with Wilms tumors and healthy patients. This is based on immunoscreening patient serum against an antigen collection (using the SEREX method) while a number of bioinformatics tools and databases are used to characterize the antigens for which a significant association with tumors was found.

SC3: Correlating Phenotypical and Genotypical Profiles (ICGKD)

This scenario, based on published microarray data, aims at finding genomic signatures for various patient phenotypic data. It involves uploading data obtained from a publicly available (microarray) dataset and the association of clinical data to the samples.

SC4: Reporting of Adeverse Events and Sever Adverse Reactions

This scenario a proposal for using the ACGT infrastructure as a reporting tool for adverse events occurring during trials. Specialized statistics might be done on the collection of adverse event cases.

SC5: In-Silico Modelling of Tumor Response to Therapy

This scenario aims at predicting the evolution of tumors based on computer models, while validation will be conducted based on the imagind data provided to the ACGT infrastructure.

SC6: Molecular Apocrine Breast Cancer

This scenario, based on published Affymetrix-array data, illustrates the application of bioinformatics tools in clinical-trial context to identify a new subcategory of breast cancer.

SC7: NKI2 Study

This scenario, based on published oligo-array data, illustrates the validation of a 72-genes breast-cancer prognostic signature.

SC8:Antigen Characterization

This scenario is a refinement of the antigen characterization section of Scenario SC2 which incorporates usage of text mining tools.

Not all those scenarios relate directly to the tools developed in the context of WP6, and some of them have some redundancy in the tools they are using. In the context of WP6 a subset of this list of scenarios will be developed further and used as validation basis for the platform, namely SC3, SC1, SC2 and SC6. Section 5 describes those scenarios in greater detail.

3.2 Knowledge Discovery in the ACGT infrastructure

Knowledge Discovery (KD) can be described as the task of extracting novel, valid, potentially useful and interesting knowledge from data. The ultimate goal of the ACGT data analysis environment is to support the user in extracting new knowledge about cancer from clinico-genomic data and to enable him to easily share knowledge about both cancer and scientific analysis of cancer in a Grid environment. This means that the ACGT data analysis infrastructure has to support three types of activities:

- **Analysis of clinico-genomic data with standard techniques**. This includes a user-friendly access to these analysis operators, the orchestration of a large set of these operators in a complex analysis workflow, but also support in choosing and evaluating operators for specific tasks and interpreting their results.
- **Development of novel analysis techniques**. Putting a new analysis algorithm to work in the scientific analysis of clinico-genomic data can be a demanding tasks, because of various practical problems, from getting access to the right data over executing the right pre- and post-processing steps to make the analysis meaningful up to problems of finding the right computational resources. Hence, a Grid-enabled analysis platform should support the development and practical deployment of novel algorithms by facilitating their implementation and integration in existing workflows
- **Sharing of knowledge**. One of the major goals of the ACGT platform in general is to ease the access to various information sources in the field of cancer research, including databases from clinical studies, public databases, and published literature. However, to actually find the relevant information for a given task from this vast pool of data sources is not a trivial task. At the same time, new knowledge about cancer will be generated by the ACGT platform in the form of novel analysis workflow and well-validated results. Supporting the user in structuring and interpreting this knowledge and finding the relevant pieces of information will be crucial for the practical success of the platform.

The data analysis tools have to be integrated into the global ACGT infrastructure, which has a complex layered organization, involving security, data anonymization GRID-distributed computing, database access uniformization, GUI design, etc. Schematically the KD tools are represented by the pink box in Figure 1. This figure shows a simplified representation of the ACGT infrastructure in which only the components of direct relevance for WP6 are depicted.

The end-user interface to the ACGT infrastructure is the Portal (which can be a browser-based application or a command line interface). The data processing requests of the user are sent to the KD “engines” which in turn request data from databases through the mediator. The latter will guarantee a uniform access to database, notably at the semantic level by using ACGT ontologies.

Scenarios based on ACGT trials and on published data (technology-driven scenarios) will be used as test bed for the WP6 tools.

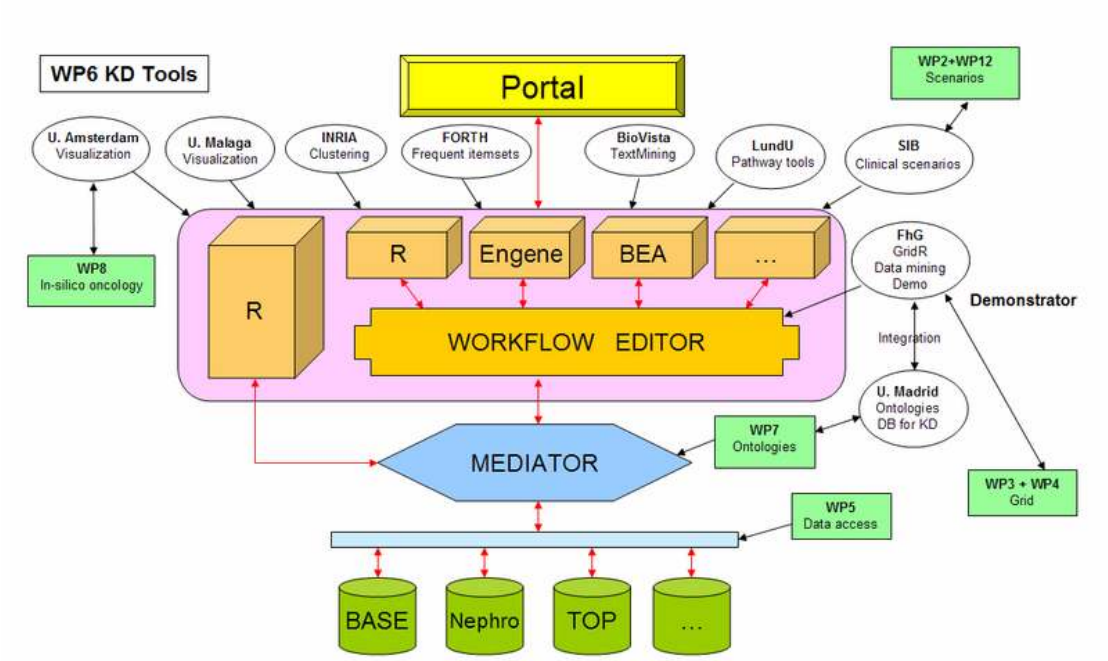


Figure 1: Overview of WP6 tools organization (pink box), interfaces with other major ACGT components (red arrows) and contributions from WP6 partners (ellipses).

3.3 Overview of interactions between KD components

Depending on the user needs, two “engines” will be available in the tools proposed in WP6. On one hand a workflow editor will let biomedical researchers combine standard software components in the KD toolbox; the workflow editor will also be used to offer predefined data analysis paths to clinicians and biomedical researchers. On the other hand, the statistical software package R will be able to access the Grid services directly, thus will offer a command-line based access to the ACGT environment. This feature is of great importance for data miners who are willing to develop new algorithms for data mining. It should be noted that R will play a dual role as it can be a workflow component too, thus providing access to a large collection of pre-existing clinical data analysis tools.

In general the software tools developed in the context of WP6 can be seen as “plugins” for the workflow editor, with the contributions of the various partners in WP6 sketched in the ellipses in Figure 1. As KD tools will in general be linked in a chain, the input and output of each of those needs to be specified. For instance, after loading microarray data from a database and having visualized them, the user may apply a filter to select a subset of those data (e.g. patients with a given phenotype). Then after clustering on the genes in the microarrays, the user may be interested in finding the pathways which involves those genes and retrieve information about them. Such scenarios depend critically on a proper propagation of information across tool interfaces. Besides describing tools on their own, it is one of the purposes of the present document to facilitate the definition of such interfaces.

2 Global Architecture

2.4 Grid-based Data Mining

Knowledge Discovery in Databases (KDD) is a research field located at the intersection of Machine Learning, Statistics, and Database Theory, and is often

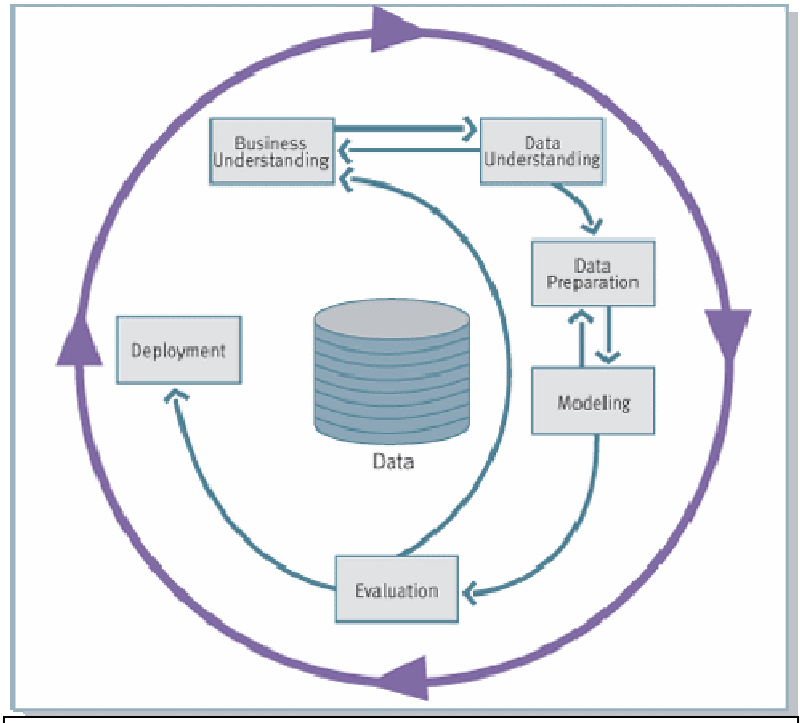


Figure 1: the CRISP Data Mining Process Model

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:

- **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.
- **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.
- **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.
- **Modelling:** the actual analysis of the data using algorithms from Machine Learning or Statistics.
- **Evaluation:** checking the correctness and applicability of the model in the application context with respect to the goals of the analysis task.

- **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 R (http://www.R-project.org) or Weka (http://www.cs.waikato.ac.nz/~ml/weka/) 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].

2.1.1 Data Mininig in Biomedical Data

Data mining methodology and technology has been developed for classical business, finance, and customer-oriented application domains. Such domains are characterized by the availability of large quantities of data in an attribute-value based representation, high ratio of examples over attributes in the data set, and weak background knowledge about the underlying entities and processes.

For biomedical data these conditions do not hold. Although technologies like microarrays for gene expression profiling are rapidly developing, today it still remains an expensive technology. In addition, legal, ethical and practical limitations in clinical trials make it cumbersome to acquire a high number of patients in a clinical trial. As a result, a typical genomic data set may contain only about 100 examples. At the same time, the same data sets consist of more than 10^4 attributes (genes). Under these conditions, standard statistical and machine learning methods are likely to over-fit the structures in the data, such that a high amount of domain knowledge is needed to guide the analysis and guarantee the validity of the extracted knowledge.

A specific property of the biomedical domain that make it very challenging for KD is its heterogeneity, both in terms of data and in terms of use cases. Concerning the data, next to genomic information very different forms of data, such as classical clinical information (diagnoses, treatments) and imaging data (x-rays, CTs) have to be integrated into the analysis. Additionally, most of the high-level knowledge is present in electronic texts, such as journal papers, which can be exploited by methods of text mining. Likewise, use cases can differ very much because of the different user groups involved. There are at least three users groups, the clinicians, who want to treat single patients, biomedical researchers which want to acquire new knowledge about genes, and data miners, which are interested in the analysis algorithms per se. All these groups have different interests and very different expertise and views on the same problem. A fruitful collaboration requires that it is easy for each user to benefit from the knowledge of the other user groups without needing to become an expert himself.

2.1.2 Mining Biomedical Data on the Grid

Grid computing is a generic enabling technology for distributed computing. It is based on a hardware and software infrastructure that provides dependable, consistent, pervasive and inexpensive access to computing resources anywhere and anytime. In their basic form, these resources provide raw compute power and massive storage capacity. These two grid dimensions were originally dubbed Computational Grid and Data Grid, respectively. Standard grid solutions provide services such as job execution and monitoring, parallelization, distributed data access and security.

The combination of data mining and grid technology offers many interesting scenarios for scaling up data mining tasks and approaching tasks not possible before in stand-alone or cluster environments:

- **Distributing data**: the integration of relevant, heterogeneous, possibly steadily updated information from distributed sites is practically a very difficult task without adequate standardization.
- **Distributing computation**: end users often may not have large computational resources at their hands, but may need to rely on other high-performance computing facilities made available to them.
- **Flexible combination of both**: a particular property of knowledge discovery is that the input data is typically not analyzed in it's raw form, but several pre-processing steps are needed. Applying these pre-processing steps directly at the sites where the data is located can result in a massive reduction of the size of the data that needs to be transported.
- **Parallel computation**: In the majority of cases, KD algorithms consist of a set of almost identical, independent tasks, *e.g.* for parameter sweeps, or feature selection. These are trivially parallelizable and can be executed on available low-cost processors. For example, in many organizations, such as a large hospital, the administrative computers are idle during the night and could be integrated for in-house data mining tasks.
- **Security**: Protecting the security of the individual data sources, e.g. for privacy-preserving data-mining.

All these tasks by themselves are not complicated to implement, as appropriate techniques have been well-known for years. The important new contribution of grid technology is to provide a standard architecture that guarantees the correct execution of the jobs, the consistency of the data, and the easy delivery of data and algorithms across different sites.

2.1.3 Ontology-enabled Knowledge Discovery

Ontology-enabled knowledge discovery operations are to be supported on the basis of standard meta-data annotation schemas for:

- ontology-based description of the application domain- taking into account standard clinical and genomic ontologies, nomenclatures and metadata (as presented before), in order to retain semantics on all steps of the analysis and to guide the construction of data mining workflows; and
- description of knowledge discovery tasks and operators (e.g., clustering, feature-selection, classification, discovery of frequent itemsets, visulaization etc), including support to correctly translate and align biomedical research

questions (in the context of clinico-genomic trials) into specific data mining tasks. Here, standards for annotating data-mining services and tasks are to be utilised into two directions:

- process standards, like CRISP-DM [CHA1999], an industry, tool and application neutral standard for defining and validating data mining processes, and
- model defining standards, like PMML (Predictive Model Markup Language)- an XML language which provides a way for applications to define and exchange statistical and data mining models by describing the inputs to data mining models, the transformations used prior to prepare data for data mining, and the parameters which define the models themselves [PMML2005].

2.1.4 References

[CHA1999]	C. Chapman et al., "CRISP-DM 1.0", http://www.crisp-dm.org.
[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.
[MIT1997]	Mitchell, Tom (1997), "Machine Learning", McGraw Hill.
[PMML2005]	PMML Schema Version 3.1: http://www.dmg.org/v2-1/GeneralStructure .html.

3.5 User Interfaces

Knowledge discovery is an interactive, iterative task. In particular, the requirements of novelty and interestingness of the discovered knowledge call for a close interaction of the knowledge discovery algorithms and the users, who have to assess the relevance of the newly found information. A trial-and-error-based working style with frequent changes in the assessment of interestingness of results and appropriateness of solutions that come with new insights is customary in this field. Hence, an interactive, flexible user interface is crucial for the success of any knowledge discovery environment.

At least three different user types that are supposed to work with the ACGT data analysis environment can be identified.

- End users (e.g. clinicians), which have little expertise in statistical data analysis and are mainly interested in running pre-defined analysis workflows.
- Biomedical researchers, who experienced in standard statistical and data mining techniques, and are interested in acquiring novel knowledge in the bio-medical domain.
- Data miners and statisticians, who are not interested in the bio-medical domain per se, but in the construction and evaluation of new data analysis algorithms.

The requirements of the user groups in terms of usability and flexibility of the user interface can hardly be integrated into a single user interface. Instead, specialized user interfaces for each user group have to be developed.

2.1.5 Portal

The ACGT Portal, developed by WP 14, will be the main entry point to the ACGT environment by new users. The goal of the portal is to provide an easy web-based access to the ACGT platform, plus training facilities to help to user to explore the full potential of ACGT. As the knowledge discovery capabilities are an important part of ACGT, this includes that new users should be able to easily perform data analysis in the portal. On the other hand, it is not practical to include the full potential and flexibility of a dedicated data mining environment into the portal; at the same time it is assumed that users on the portal are more interested in getting an overview and introduction to ACGT instead of performing detailed analyses. Hence, it seems reasonable to require that the portal should provide the means for the user to execute pre-defined analysis workflows on his own data set, while the construction of new workflows and new operators is left for a more suitable user interface.

An exemplary user session on the portal could be as follows:

1. The user logs into the portal.
2. The user uploads his data and describes its semantics in terms of the ontology.
3. The user chooses an analysis tasks to perform, which corresponds to choosing a pre-defined workflow.
4. The workflow is executed on the user's data set, prompting the user to input parameters when necessary.
5. The results are displayed into the portal, or can be downloaded from the portal.

Several technical requirements follow from this scenario. For example, to exploit the full potential of the semantics provided by the ontology, it must be possible to define workflow steps in terms of semantic concepts instead of the specific format of the data, meaning that for example it must be possible to formulate a workflow step in the form of "apply operator O on all clinical attributes of the data set" instead of actually listing up all attributes. This allows to define generic workflows, which are not specific to a single data set or data format.

It must be noted that only a sub-class of workflows may be executed in the portal. In particular, workflows which require the user to download and install software on his own computer, e.g. for advanced interactive visualization, would better be executed in the workflow editor. For the sake of usability and portability, results of workflows in the portal must be limited to data types that can be presented in a typical web browser, e.g. as text or images.

2.1.6 Workflow Editor

The workflow editor, which will be developed by work package 9, will be the central user interface of ACGT. A workflow editor is a graphical interface for the creation and execution of workflows, and thus provides the means to flexibly connect single operators and services into a meaningful algorithm without the need to actually learn a programming language. Another main task of the workflow editor is to check the

syntactic correctness of a workflow prior to execution, in order to avoid problems that might occur at runtime.

This requires that the workflow editor must supply a detailed set of interfaces and data types in order to express the objects that can occur in a data mining workflow. Necessary data types include

- *Data Sets*: data sets contain the actual clinico-genomic data that is to be analyzed within ACGT. Different have to be taken into account in order to allow an easy integration of new analysis algorithms (possibly automatically invoking conversion if necessary).
  Among the possible data types, propositional data is the most most frequently used case in data mining. Propositional data means data with a fixed number of attributes to describe each data instance, e.g. using blood pressure and age to describe a patient. In order to make these attributes usable for data mining, several properties of them have to be described. Next to their semantics in terms of the application domain, which will be dealt with by the ACGT Master Ontology and the Mediator, necessary information for data analysis includes the role of each attribute in the analysis process (attribute, label, prediction, id), its data type (Boolean, nominal, numeric, integer, real, text etc.). Additionally, some statistical properties of the attributes, e.g. the number of missing values or their mean or mode, may be important when trying to decide whether a certain algorithm can work on a given data set (e.g., there are many algorithms which can only work on numerical data without missing values).
  Additional data types that may become relevant within ACGT are relational data (e.g. data coming from relational databases), and data types that are specifically used in the bio-medical domain.

- *Models*: The results of data mining algorithms which describe the extracted properties of the data are called models. Models typically make only sense in combination with the analysis algorithm that constructed these models. For example, when a classification model shall be applied to predict the class of a previously unknown observation, the classification algorithm must be used to execute the model on the new observations. Additionally, a user may be intereste in inspecting the model, where again the type of visualization to be used depends on the model.
  In summary, a model may be viewed as a black box that does not make sense without a specific algorithm. In order to increase usability and prevent avoidable errors, a generic application or visualization operator, that calls the specific application and visualization operators depending on the type of the model, should be implemented.

- *Parameters:* Most analysis and visualization operators come with a set of parameters that can be set by the user und whose settings have a large impact in the quality of the results. Hence, an important step in many data mining applications is the automatic or semi-automatic tuning of parameters. To facilitate this tasks, a set of parameters must be supported as an own data type; this enables them to can be transferred between the optimization subtask, and the predictive subtasks of a workflow and allows to return the optimal set of parameters as a data mining result.

- *Performance values:* In data mining research on predictive models, one is typically not interested in the predictions itself, but in the performance of an

algorithm for the task as a whole. In order to make comparisons between different data mining algorithms, performance values with respect to different accuracy or loss measures and support for the comparison of these values must be implemented.

- *Images and Texts*: results of data analysis can often very easily and intuitively be summarized in automatically generated images and texts. In particular with respect to the connection of the workflow editor and the portal, images and text are the types of data that can be displayed on a standard web browser without any conversion. Hence, these data types need to be supported as outputs of a workflow.

2.1.7 R

In finding novel ways to analyze such complex data as in clinico-genomic studies, there comes a point where the combination of existing techniques such as in a workflow editor is not enough and new analysis algorithms have to be developed. In order to ease the access to the data sources and services that will be supplied in the ACGT platform, a programming language interface that supports both the access to the ACGT services and is easy to use for the intended users, namely bio-statisticians, will be necessary. The open source statistical package R (http://www.R-project.org) has quickly become the platform of choice for statistical research and many applied statistics project. 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.

Hence, the R environment is the obvious choice for the programming language interface of the ACGT data analysis environment. Note that R will also be integrated as a single analysis tool that can be executed from workflow editor (see Section 4.1) and thus provides a large spectrum of statistical tools and services. However, in this section we will concentrate on the use of R as a user interface.

The integration of R into the ACGT analysis environment needs to support the following functions

- *Access to ACGT data types:* the data types that are supported by the ACGT environment, in particular the mediator and the workflow editor, must be importable into R. This includes both the access of data, but also the mapping of meta-data to a compatible representation in R.

- *Access to ACGT analysis services:* the services that are supplied by the ACGT analysis environment must be accessible from within R. In particular, this includes the execution of ACGT workflows as subtasks of an R function.
- *Access to ACGT Grid services:* The ACGT Grid services, in particular the remote execution of functions and parallelization of expensive computation tasks, should be easily usable for R code. Also the lookup of Grid services must be supported.
- *Access to ACGT administrative services:* This includes in particular the support of user management and security.

With respect to usability, no specific requirements except for a standard documentation arise in the case of the R interface, because it is supposed to be a pure programming language interface and is supposed to be used by experts only.

3 Semantic Integration

One of the main goals of ACGT is to achieve the semantic, ontology-based integration of clinical and genomic/proteomic data. With respect to knowledge discovery, this approach promises valuable new perspectives for exploiting and integrating semantics in the analysis process, and putting newly discovered knowledge into perspective regarding what is already known. Semantic knowledge has historically been left aside in KDD analysis, mainly due to the lack of support for its representation and utilization. Due to this fact, results produced by KDD analysis sometimes suffer from a lack of interest, and lots of information and features keeps undiscovered. Ontologies can contribute in this aspect.

Ontologies can be the key to solve the problems mentioned above. Ontologies and metadata allow describing areas of knowledge in a formal manner. Their integration into the knowledge discovery process can allow finding new knowledge, and supporting knowledge management for the optimal exploitation of the found knowledge. For example, an ontology can be designed to describe the domain of existing data mining algorithms, so it can be processed by a machine, allowing it to *decide* which algorithm fits better with a given task. Ontologies have been proven to be a suitable tool to perform such task. This approach will produce a decrease in time consume and more fruitful KDD results.

There are several ways or approaches that can be taken to enhance KDD by using ontologies, depending on which KDD phase is to be enhanced. The most important (and most investigated) are the following ones:

1) ***Data integration***. Ontologies can help integrate disparate sources containing heterogeneous data. Ontologies can act as a common framework of information, providing the necessary semantic background for proper integration of heterogeneous data. It must be noted that by heterogeneous, it is meant by format, not by meaning. Two data sources may contain data regarding the same area of knowledge (for example cancer), but employ different formats or identifiers. Most work is related to this area.
2) ***Data preparation*** (data cleaning, feature selection,…). Ontologies can facilitate cleaning of data, by effectively organizing the required knowledge for this task. They can also provide a solid background for the design of tools that provide advice to the engineer in the feature selection task.

3) ***KDD process selection***. A novice (or even an expert) KDD engineer can be overwhelmed by the large amount of available algorithms for KDD tasks. Ontologies can be used to store and organize the knowledge regarding KDD algorithms, allowing the design of programs to be able to analyze and provide KDD engineers with advice for the most appropriate one.
4) ***KDD result description and reusing***. Ontologies can provide a formal representation of the domain of the new knowledge obtained. It can therefore help in the reutilization of knowledge for new KDD iterations.

Most work in this area is related to the data integration approach. Examples of this are the methodology proposed by P. Gottgtroy & colleagues [GOT2003][GOT2004], a system developed by A. Silvescu & colleagues [SIL2001], the OUIP tool [CHO2003], ONTOFUSION tool [PER2005] (developed by UPM), etc.

A detailed description of the topic of ontologies and their use in the ACGT project can be found in Deliverable 7.1 *"Consolidated Requirements on ontological approaches for integration of multi-level biomedical information".*

3.1.1 Ontologies & KDD

Description of KDD

KDD (Knowledge Discovery in Databases) is the process of knowledge extraction for raw data. It appeared in the 90's and tries to solve the problem of managing very large volumes of data (obtained in different ways) and obtaining useful information (knowledge) from it.

The KDD process is composed by several steps, which were briefly described by Fayyad as i) understand the domain and goals, ii) create the data set, iii) preprocess the data, iv) select the data without loss of information, v) data mining and vi) pattern interpretation.

Problems of KDD

Very much effort has been dedicated to developing effective algorithms for data mining; however, the data preparation process (data extraction, data selection and data clearing) has been rather left aside. Data preparation is one of the steps that more time and effort requires (and thus more resources consume), mainly because most times it is manually carried out. It is therefore necessary to develop methods that allow automating (totally or partially) and enhance data preprocessing.

This is not the only problem of KDD process. The great amount of available algorithms can also make difficult for the data mining engineer (either novice or expert) to conduct successfully the whole process. It is sometimes hard to make the right choice and thus it would be desirable to partially automate this task. Results of a KDD process are often not managed as well as it could be. The extracted knowledge is often not used in subsequent analysis.

Objectives

The main objective is to solve the problems described above. Specifically data integration, data selection and data cleaning and feature selection are the steps that require most emphasis.

Ontologies to enhance KDD

Ontologies can be a solution to face these problems. With an ontology it is possible to describe the domain of some area of knowledge in a formal manner. This way a new semantic layer is added, providing more powerful possibilities to the KDD process. Cespivova & colleagues [CES2004] demonstrated that ontologies can be beneficial in nearly all phases of KDD process. They analyse the KDD cycle as described in CRISP-DM (Cross Industry Standard Process for Data Mining):

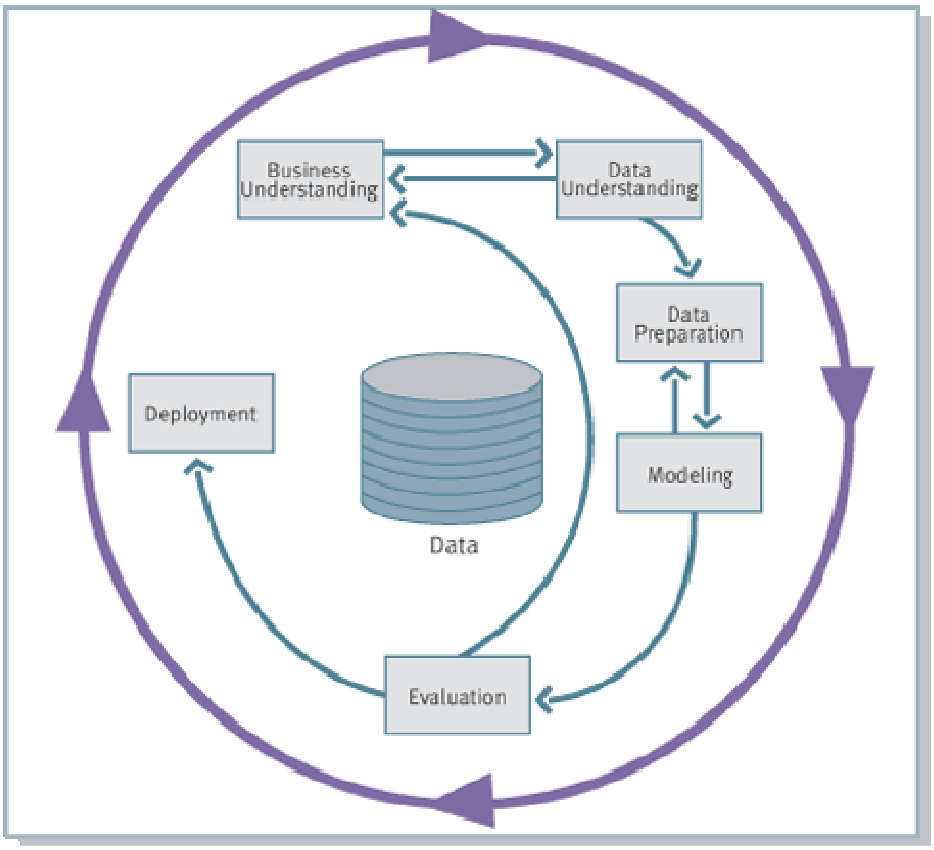


Figure 2: CRISP_DM model for KDD cycle

Six main phases are defined in this model. They are listed below, including suggestions for different roles of ontologies:

- Business understanding: domain ontologies can help a newcomer to get familiar with most important concepts and relationships.
- Data understanding: elements of the ontology must be mapped to elements of the data scheme and vice versa. An easier identification of missing and redundant attributes could be achieved.
- Data preparation: an ontology might help identifying multiple groupings for attributes and/or values.
- Modelling phase: ontologies might help designing the individual mining sessions.

- Evaluation phase: discovered models can be interpreted in terms of an ontology and the associated background knowledge.
- Deployment phase: the extracted knowledge is fed back to the business environment. If the business was previously modelled by means of ontologies, the new knowledge can be easily mediated by such business ontology.

APPROACHES

There are several ways or approaches that can be taken to enhance KDD by using ontologies, depending on which step is to be enhanced. The most important (and most investigated) are the following ones:

- Data integration. Ontologies can help integrate disparate sources containing heterogeneous data. Most work is related to this area.
- Data preparation (data cleaning, feature selection …). Ontologies can facilitate cleaning of data, or be helpful in the design of tools that provide advice to the engineer in the feature selection task.
- KDD process selection. A novice (or even an expert) KDD engineer can be overwhelmed by the large amount of available algorithms for KDD tasks. Advice for choosing one can be facilitated by tools based on ontologies.
- KDD result description and reutilization. By having a formal representation of the domain of the new knowledge obtained it should be easier to reuse that knowledge for new KDD iterations.

3.1.2 Methods & Models

Methodology proposal (P. Gottgtroy, N. Kasabov & S. MacDonell)

In these papers [GOT2003][GOT2004], a new methodology is proposed, and a framework that aims to make full use of the semantic power of ontologies in the KDD process.

They describe the Infogene Map ontology. It is a case study to build a multi-dimensional biomedical ontology. It includes the information of six different ontologies from the biomedical area.

It can be argued that the benefit of applying ontologies in the KDD process follows two different directions. Ontologies drive the KDD process from a data driven approach to a knowledge driven approach, whereas KDD can provide a useful input for ontology learning. They define these two approaches as Onto4KDD and KDD4Onto respectively. A new combination is created: Onto4KDD4Onto, which gathers the two previous terms.

Figure 41 shows the proposed framework:

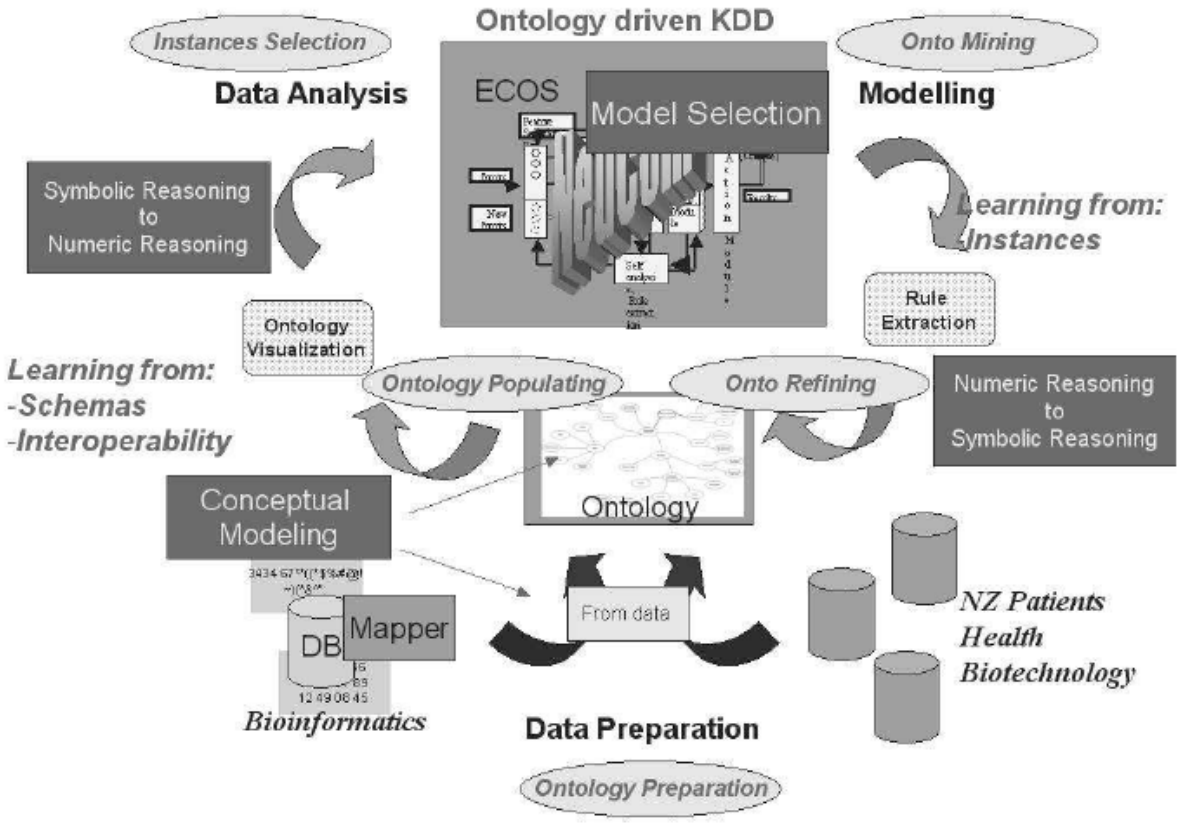


Figure 3: Onto4KDD4Onto framework

The methodology for including ontologies in the KDD process consists on five steps, which are described below:

- Ontology preparation: both the data model and ontology model must be analysed. This is necessary for proper matching of concepts.
- Ontology population: the ontology in use is expanded with concepts retrieved from different databases.
- Instance/Feature selection: this is one of the most important steps, since a good feature selection will be crucial in a successful KDD process, producing useful results.
- Ontology mining: in this step the ontology used learns from the KDD process, producing new relations between existing concepts. Thus, the ontology model is refined and updated.
- Ontology refining: in this step numeric results must be translated into concepts that can be added to the ontology.

Methodology proposal (V. Svatek, J. Rauch & M. Flek)

The paper [SVA2005] investigates the use of ontologies in the KDD process in the domain of Social Reality. The authors created an ontology that was enlarged with different concepts. These were all mapped to existing databases related to the

mentioned domain, using Lisp-Miner for this task. Their hope was to aid the KDD process by providing human understandable explanation of the obtained results.

Status: work in early phase

3.1.3 Systems & Tools

This section contains the systems and tools that use ontologies to enhance the KDD process.

IDEA

IDEA [BER2001][BER2005] is an IDA (Intelligent Discovery Assistant) prototype. It is therefore intended to support KDD researchers in their work. Given a task, IDEA provides with enumerations of suitable KDD process, ranking them by several criteria.

IDEA is based on an ontology containing knowledge about KDD processes (and their inherent characteristics). The tool works like follows:

- Characteristics of the data to be analyzed, and the expected results are asked.
- IDEA explores the space of KDD processes, selecting those that match the requirements.
- It presents an enumeration of such processes, according to different criteria.

Figure 42 describes the basic structure of IDEA:

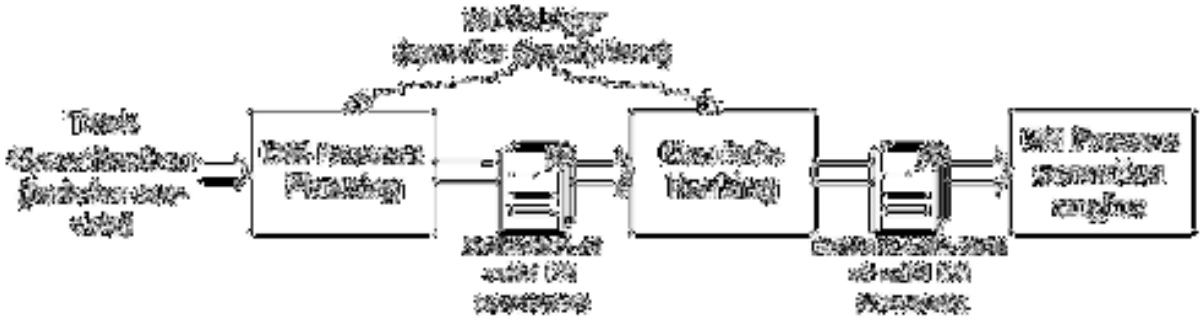


Figure 4: Basic structure of the IDEA project

IDEAs first core component is the DM-process planner. It uses an ontology to assist the user in composing valid and useful DM processes. This ontology contains for each operator:

- The conditions under which the operator can be applied.
- The operator's effects on the DM state and the data.
- An estimation of the operator's effects on attributes such as speed, accuracy, etc.
- Comprehensible information about the operator.

IDEA is composed of two core components: the DM-process planner and the heuristic ranker. The former is in charge of navigating through the ontology, searching for suitable KDD processes, given a problem, while the latter must rank the processes selected previously, based on criteria such as speed, accuracy, etc.

According to the authors, several experiments have shown promising results, helping researchers in their data mining work.

A system for data integration from heterogeneous sources (A. Silvescu & colleagues)

This work [SIL2001] aimed to address some of the challenges that integration and analysis of heterogeneous and distributed data presents (such as the difficulty of accessing the sources in a homogeneous way). Ontologies are presented to store an important background knowledge which can help in the KDD process providing the necessary context. Figure 44 shows simple example architecture of ontology-based KDD:

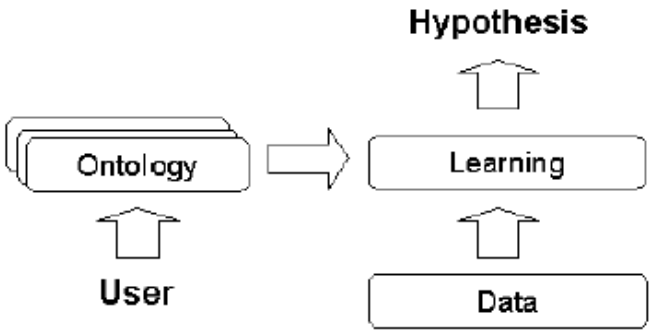


Figure 5: Ontology-based KDD architecture example

A prototype system for integration of distributed data is presented. It can automate all the information extraction and transformations, and has been successfully tested with Swissprot and Prosite. Data integration results in its storage in a central warehouse. This process is accomplished by means of an ontology which specifies the relevant relationships among entities in our universe of discourse.

One main limitation of this system is that the system uses default ontologies, and no others can be specified. However this is planned to change in the future, allowing for custom ontologies to be used. Figure 45 shows the system architecture.

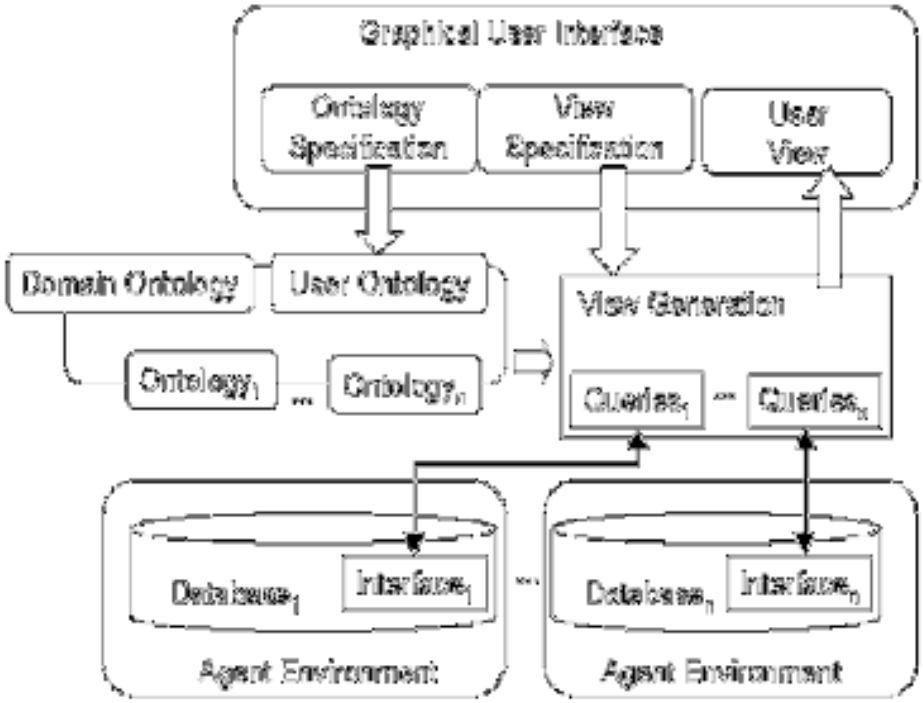


Figure 6: A snapshot of the system architecture

Current status: unknown.

System proposal (J. Phillips & B. G. Buchanan)

This paper [PHI2001] describes a proposal for a system capable of suggesting feature selection in the KDD process, therefore reducing user workload. To do this, knowledge stored in ontologies is used. Databases were discarded because they lack the required semantic knowledge.

Their system scans databases to obtain type and constraint information, which users verify. The knowledge of a shared ontology is then used, allowing the system to intelligently guide the feature construction process. As a result, the KDD process is improved, and user's workload is significantly reduced. Figure 46 shows a snapshot of this process.

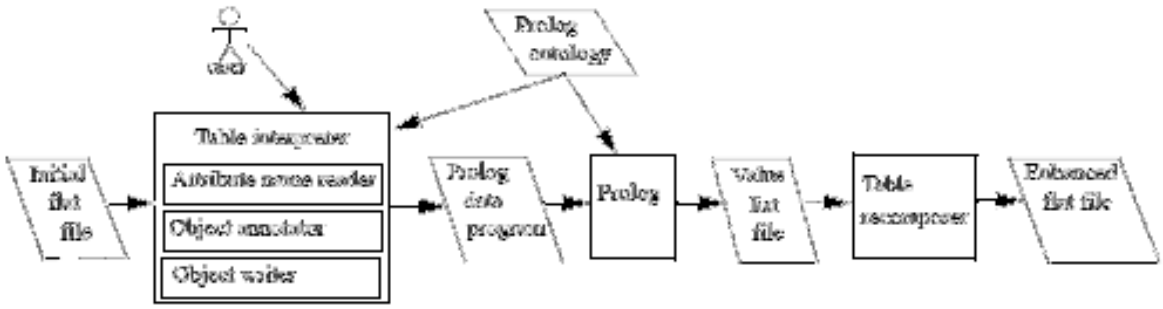


Figure 7: Overview of the process that takes in place in the system

The system is composed of three components, which act sequentially to compose the overall process:

- Table interpreter: in charge of the creation of attribute domains. It uses flat files as input, and its results must be verified by the user.
- Prolog program: this program must invent new attributes, using the ontology for this task.
- Table recomposer: the newly created attributes are incorporated to the existing ones, conforming a new flat file

Moreover, the system is capable of learning each time it is used, therefore becoming more useful over the time.

LinkFactory

This paper [VER2003] presents the LinkFactory tool, already described in section 3.2.5.2, and which allows integrating biological data. In order to do this, the LinkBase ontology was constructed, containing knowledge from the biological domain. Concepts from Gene Ontology and other databases (all belonging to the molecular biology domain) were incorporated to LinkBase, hence extending the reach of the ontology. Data from Swiss-Prot was mapped to this ontology.

LinkFactory was developed as a tool that provides users with an easy to use but also powerful interface to interact and manage large and complex ontologies, such as LinkBase. It is composed by a multiple windows environment, and based on beans (more than 20). It provides a wide range of functions, as well as useful views of ontologies. Figure 47 shows the tool aspect.

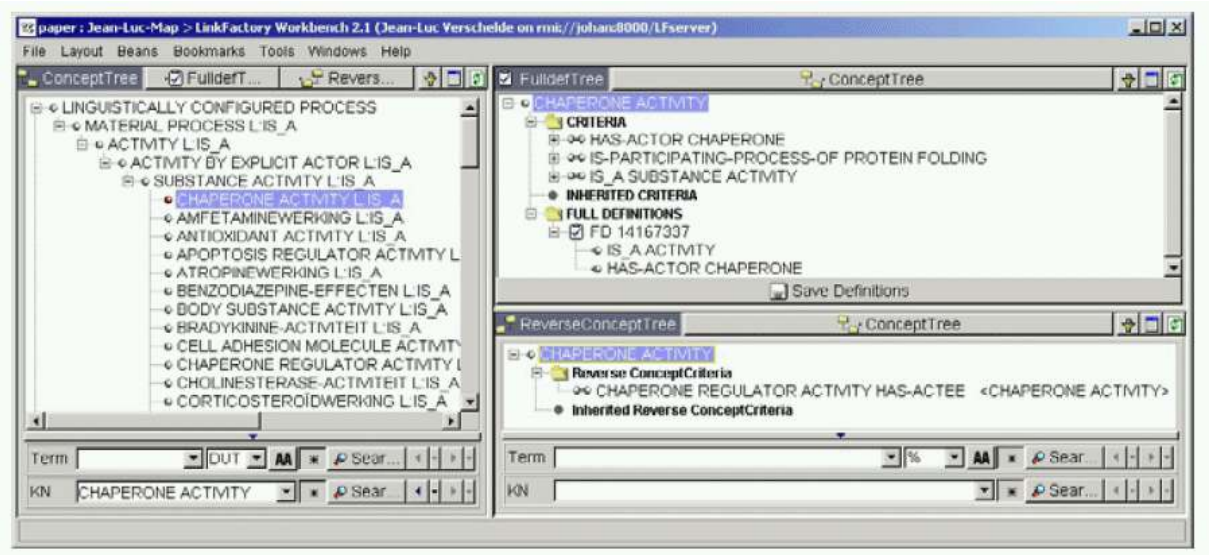


Figure 8: Snapshot of the LinkFactory tool

MiningMart

These papers [EUL2004][MOR2003] describe the application of ontologies in the KDD preprocessing tool MiningMart. This tool provides a graphical environment that allows sharing knowledge about successful KDD results. The researcher is assisted for choosing the most appropriate representation of data and selecting the more suitable KDD algorithm.

This tool is composed of a data model and a case model. The data model has two levels: the database schema and an ontology that allows describing the data in abstract terms. The case model describes operations (also by means of an ontology) that can be executed on the data (by composing such operators we get *chains*). Figure 48 shows an overview of the system.

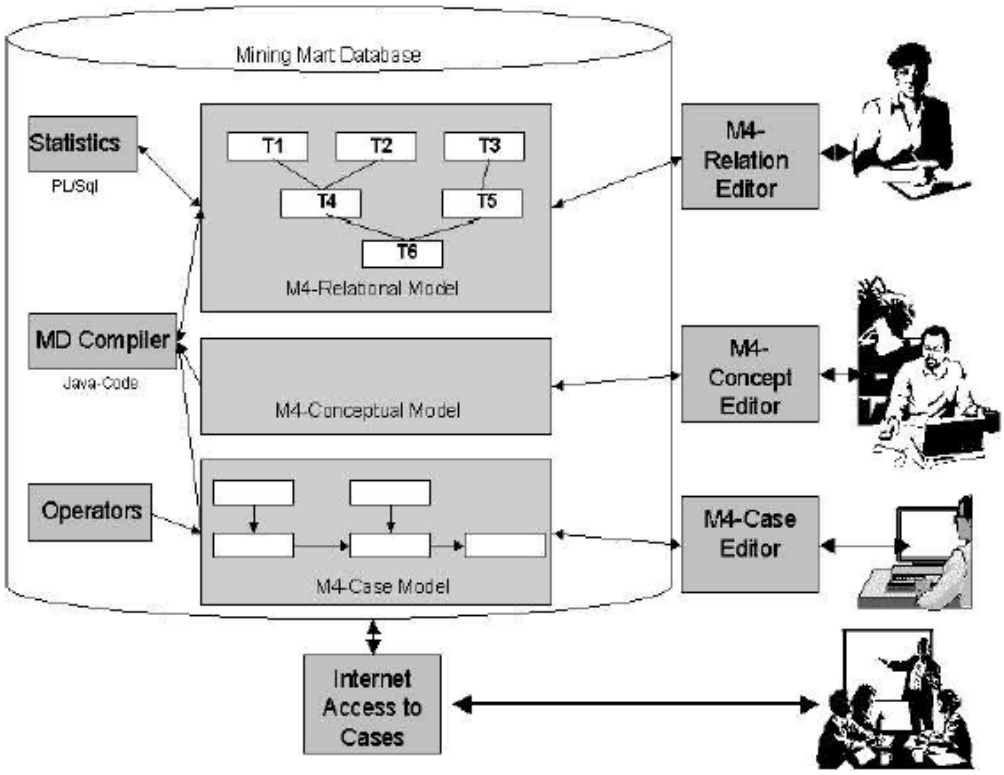


Figure 9: Overview of MiningMart system

It is stated that by using an ontology, a higher level of data abstraction is achieved. Thus, it increases the understability and reusability of the KDD process. In addition, several advantages are obtained, such as a better description of the data, automatic documentation, the possibility of reusing KDD applications and sharing knowledge about successful KDD applications.

OUIP

This paper [CHO2003] presents the OUIP tool, which is designed for allowing transparent data integration from heterogeneous data sources. A medical vocabulary is used as a repository of concepts, and data from different data sources are mapped to this vocabulary. Figure 49 shows a description of the system architecture.

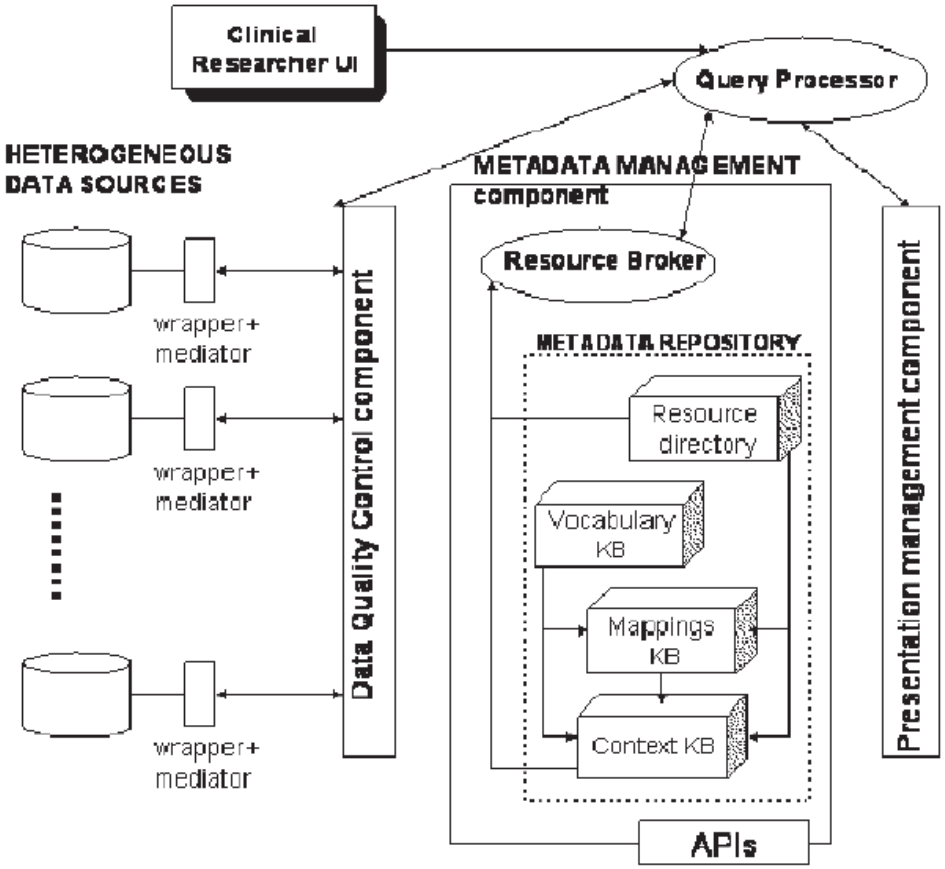


Figure 10: architecture of the OUIP system

The main components of the system are briefly described below:

- The resource broker receives user queries and selects the most appropriate data sources to answer them. To do this, it navigates in the Resource Directory looking for the appropriate metadata.

- The metadata repository stores the metadata used by the resource broker.

- The resource directory contains metadata which describes the information stored in each data source.

- The vocabulary knowledge base contains ontologies defining concepts related to the area of knowledge of the researchers that use this tool.

- The mappings knowledge base stores the mappings between data sources and concepts in the ontologies.

ONTOFUSION

ONTOFUSION [PER2005], already mentioned before in section 3.2.5.5, is an ontology-based system for biomedical database integration. It uses two processes: mapping and unification. Mapping process uses ontologies to create the necessary links of database schemas with conceptual frameworks (virtual schemas) that represent the structure of the information contained in the databases. Unification process integrates two or more virtual schemas into more global virtual schema (thus

obtaining the desired integration process). On the top, a virtual schema represents all the database schemas to be integrated. The user will see this virtual schema as a database she can query, and the underlying structure of virtual schemas and databases will not be visible. Figure 50 shows an example of how data integration process works.

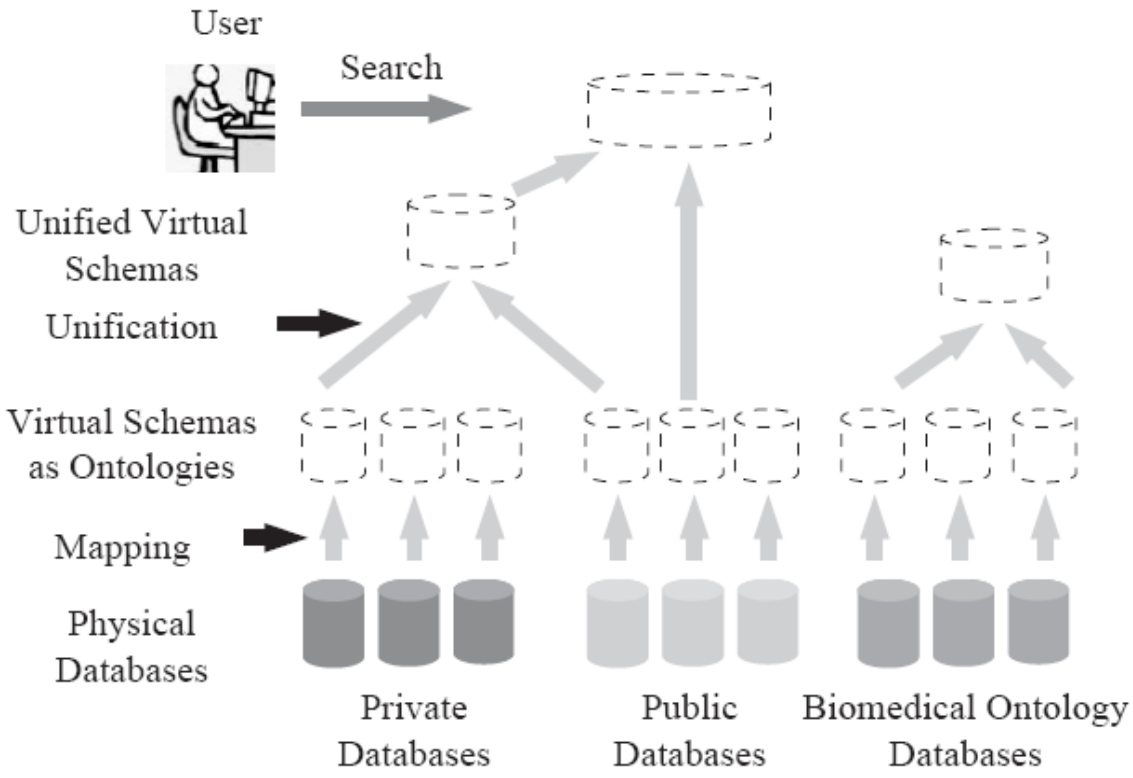


Figure 11: schema of the data integration process

The first step is to create the virtual schema (build the ontologies). Specific domain ontologies are used for this purpose, guiding the administrator in her labour. The use of ontologies guarantees that all the semantically equivalent elements in different schemas use the same concept name. Concept names cannot be chosen arbitrarily, but must be selected from the respective domain ontology.

The system can automatically unify virtual schemas that correspond to the same domain ontology.

A GUI is provided to the user so she can navigate through the different ontologies and select the database they want to browse. After this she can select concepts and specify queries.

This system began with the European project INFOGENMED, aimed to develop methods and tools for database integration from remote sources.

OntoDataClean

OntoDataClean tool, already described in section 3.2.5.5, is intended to make a simpler cleaning of data and integration from heterogeneous data sources. It is being developed by the Biomedical Informatics Group at the UPM, the group that leads

WP7. It makes use of an ontology to describe the domain of possible transformations that can be carried out on data. The user will make instances of this ontology to define how data must be transformed, increasing his understanding on what he is doing.

The tool incorporates a module for semi-automatic detection of inconsistencies that give some hints of possible transformations the data may need, reducing user's workload.

The whole system is deployed as a web service for easier access for any machine connected to the internet. It was described in more extension in a previous subsection.

Status: in progress, current ongoing work is on integration with ONTOFUSION system.

3.1.4 Platforms & Frameworks

XML framework proposal for KDD (P. Kotasek & J. Zendulka)

This paper [KOT2000] proposes an XML-based framework for the KDD process. This framework includes the use of ontologies describing the different domains of knowledge utilized in such KDD processes. This is aimed at addressing the following problems in KDD:

- Lack of precise definitions of basic concepts used in KDD (knowledge, pattern …).

- User must deal with huge amounts of data.

- Lack of formal description of available data mining tasks and processes.

- Lack of formal definition of different kinds of results.

- Inability to easily integrate new discovered knowledge with existing knowledge.

These problems should be addressed by means of ontologies, since they allow formal representation of knowledge. An ontology could be used to describe the domain of knowledge to be analyzed, allowing easier navigation in the information. Another ontology could describe the characteristics of the KDD process itself.

All this information could be integrated in an ontological library, containing domain ontologies, support ontologies and KDD ontologies. The Knowledge Sharing Effort (KSE) is already implementing this idea [KSE]. Figure 51 shows this idea.

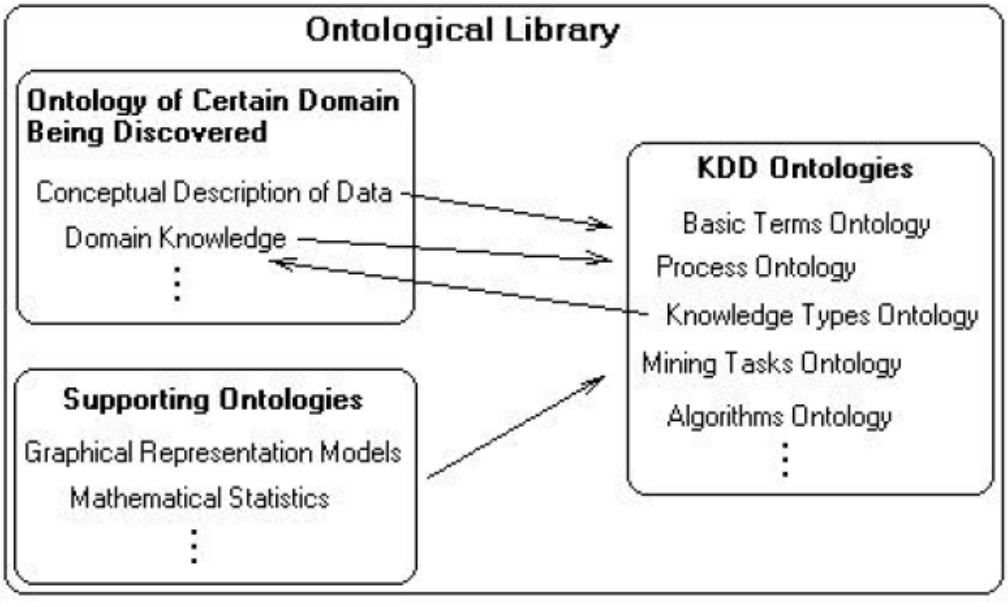


Figure 12: Possible ontological library structure

The authors recognise that they still do not have an ontology related to the KDD domain, and they propose using the XML language for its construction. More specifically, they propose the inclusion of XML Data Interfaces in the KDD process, so that each step in the process has its input and output XML data interfaces, allowing easily combining different discovery components to perform the whole process. Figure 52 shows this process in detail.

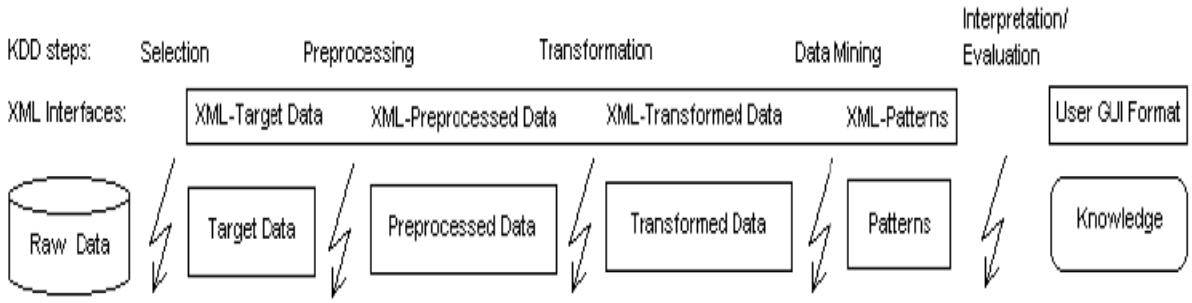


Figure 13: final KDD process, with XML interfaces incorporated

KTA framework

KTA (integrating expert Knowledge in Transcriptome Analysis) [BRI2004] is a framework aimed at improving the mining process by adding the knowledge contained in ontologies prior to the KDD process itself. KTA is part of the MEDIANTE project.

The patterns found by a KDD algorithm might be of use for a specific user, but not for a different one. That is why the authors propose including subjective measures as criteria in the mining process. Concretely, they define two criteria: unexpectedness and actionability. The former refers to the unexpectedness of a result, while the latter measures the advantage the user can take of the result.

OntoClean framework

This paper proposes OntoClean [WAN2005], an ontology-based data cleaning framework. In OntoClean, a set of ontologies describes the data domains to be cleaned, allowing the resolution of semantic inconsistencies. A template composed of five steps is provided, describing the process of data cleaning within Ontoclean. Figure 53 shows this framework graphically.

The main advantages are: the possibility of admitting input from users in natural language (hiding the details of the algorithms employed), a higher quality in the cleaning process, achieved by the use of knowledge contained in both domain ontologies and task ontologies, and the easiness of extending the system given by the use of OWL (the standard web ontology language).

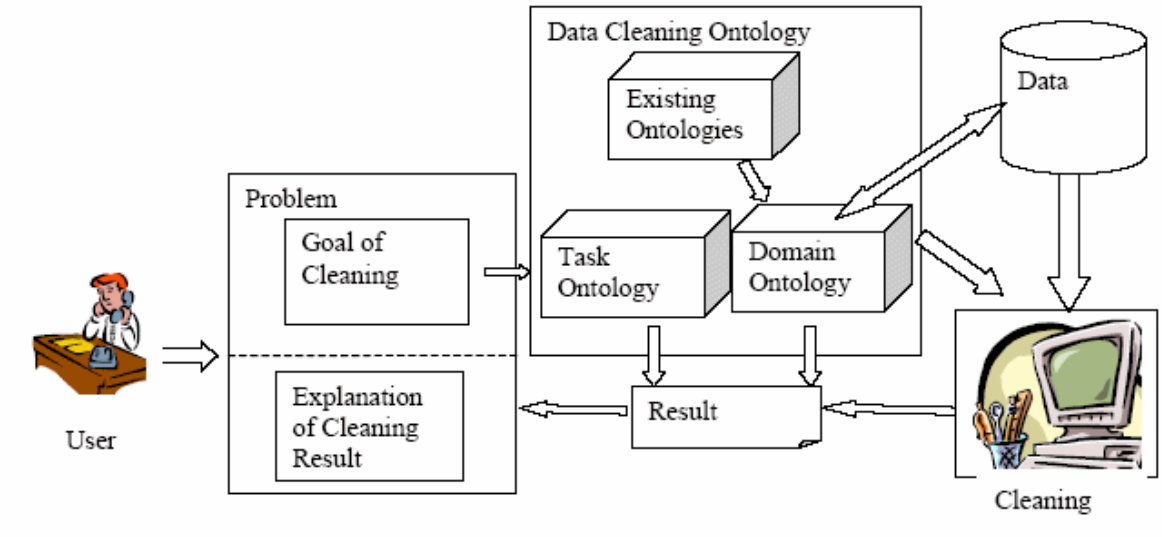


Figure 14: components of OntoClean framework

Projects

This section contains the projects related with application of ontologies to the KDD process.

DataFoundry

DataFoundry [CRI2001][DATAF] project, already mentioned in section 3.2.5.1 of this deliverable, aimed to improve scientists' interactions with large data sets. Some of the efforts were focused in improving the access to distributed, heterogeneous data. An architecture which used ontologies was developed, allowing easier data warehouse maintenance. The use of the ontology allowed automatic generation of mediators, thus making it unnecessary to manually define how data must be transformed from the data sources to the central warehouse. The metadata contained in the ontology allows solving both syntactic and semantic conflicts between data contained in the sources.

Four different concepts are represented in the ontology: abstractions (abstractions of domain specific concepts), databases (database descriptions), mappings (mappings between a database and an abstraction) and transformations (functions to resolve representation conflicts).

Status: inactive.

3.1.5 References

[BER2001] Bernstein, A., and Provost, F.: "An Intelligent Assistant for Knowledge Discovery Process" IJCAI 2001 Workshop on Wrappers for Performance Enhancement in KDD, (2001).

[BER2005] Bernstein, A., Hill, S., and Provost, F.: "Toward Intelligent Assistance for a Data Mining Process: An Ontology-Based Approach for Cost-Sensitive Classification". IEEE Transactions on Knowledge and Data Engineering, 17(4):503–518, 2005.

[BRI2004] Brisson, L., Collard, M., Le Brigant, K., and Barbry, P.: "KTA: A Framework for

Integrating Expert Knowledge and Experiment Memory in Transcriptome Analysis" In: ECML/PKDD04 Workshop on Knowledge Discovery and Ontologies (KDO'04), Pisa (2004).

[CES2004] Cespivova, H., Rauch, J., Svatek V., Kejkula M., Tomeckova M.: "Roles of Medical Ontology in Association Mining CRISP-DM Cycle" In: ECML/PKDD04 Workshop on Knowledge Discovery and Ontologies (KDO'04), Pisa (2004).

[CHO2003] Chong, Q., Marwadi, A., Supekar, K., and Lee, Y.: "Ontology Based Metadata Management in Medical Domains" Journal of Research and Practice in Information Technology (JRPIT), 35(2), pp. 139 – 154 (2003).

[CRI2001] Critchlow, T., Musick, R., and Slezak, T.: "Experiences Applying Meta-Data to Bioinformatics" In I*nformation Sciences*, v.139 (1-2), Elsevier Science Inc., Nov. 2001.

[DATAF] www.llnl.gov/casc/datafoundry/

[EUL2004] Euler, T., and Scholz, M.: "Using Ontologies in a KDD Workbench" In Workshop on Knowledge Discovery and Ontologies at ECML/PKDD, 2004.

[FAY1996] Fayyad, U., Shapiro, G. & Smyth, P.: "From Data Mining to Knowledge Discovery in databases". AI Magazine 17, 37-54, 1996.

[GOT2003] Gottgtroy, P., Kasabov, N. & MacDonell, S.: "An ontology engineering approach for Knowledge Discovery from data in evolving domains" Proceedings of the Data mining IV. Southampton, Boston: WIT press (2003).

[GOT2004] Gottgtroy, P., Kasabov, N., MacDonell, S.: "An ontology driven approach for discovery in Biomedicine" (2004).

[KOT2000] Kotasek, P., and Zendulka, J.: *"An XML Framework Proposal for Knowledge Discovery in Databases",* In: The Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases, Workshop Proceedings Knowledge Management: Theory and Applications, Ecole Polytechnique de Nantes, Lyon, France, p. 143-156 (2000).

[KSE] www.ksl.stanford.edu/knowledge-sharing/papers/kse-overview.html

[MOR2003] Morik, K., and Scholz, M.: "The MiningMart Approach to Knowledge Discovery in Databases" In Ning Zhong and Jiming Liu, editors, Intelligent Technologies for Information Analysis. Springer, to appear (2003).

[PER2005] Pérez-Rey, D., Maojo, V., García-Remesal, M., Alonso-Calvo, R., Billhardt, H., Martín-Sánchez, F., and Sousa, A.: "ONTOFUSION: Ontology-Based Integration of Genomic and Clinical Databases" En prensa para su publicación en el número especial sobre Ontologías Médicas de la revista 'Computers in Biology & Medicine', 2005.

[PHI2001] Phillips, J., Buchanan, B.G.: "Ontology-guided knowledge discovery in databases" International Conf. Knowledge Capture Victoria, Canada, 2001.

[SIL2001] Silvescu, A., Reinoso-Castillo, J., Andorf, C., Honavar, V., and Dobbs, D.:"Ontology-Driven Information Extraction and Knowledge Acquisition from Heterogeneous, Distributed Biological Data Sources" In: Proceedings of the IJCAI-2001 Workshop on Knowledge Discovery from Heterogeneous, Distributed, Autonomous, Dynamic Data and Knowledge Sources (2001).

[SVA2005] Svatek, V., Rauch, J., and Flek, M.: "Ontology-Based Explanation of Discovered Associations in the Domain of Social Reality" (2005).

[VER2003] Verschelde, J.L., 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).

[WAN2005] Wang, X., Hamilton, H.J., Bither, Y.: "An Ontology-based Approach to Data Cleaning" (2005).

3.1.6 References

[CHO2003]	Chong, Q., Marwadi, A., Supekar, K., and Lee, Y.: "Ontology Based Metadata Management in Medical Domains" Journal of Research and Practice in Information Technology (JRPIT), 35(2), pp. 139 – 154 (2003).
[GOT2003]	Gottgtroy, P., Kasabov, N. & MacDonell, S.: "An ontology engineering approach for Knowledge Discovery from data in evolving domains" Proceedings of the Data mining IV. Southampton, Boston: WIT press (2003).
[GOT2004]	Gottgtroy, P., Kasabov, N., MacDonell, S.: "An ontology driven approach for discovery in Biomedicine" (2004).
[KIM1996]	Kimball, R.: The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses, John Wiley. 1996
[PER2004]	Perez-Rey, D., Maojo, V., Garcia-Remesal, M., Alonso-Calvo, R.: "Biomedical Ontologies in Post-Genomic Information Systems," bibe, p. 207, Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'04), 2004.
[PER2005]	Pérez-Rey, D., Maojo, V., García-Remesal, M., Alonso-Calvo, R., Billhardt, H., Martín-Sánchez, F., and Sousa, A.: "ONTOFUSION: Ontology-Based Integration of Genomic and Clinical Databases" En prensa para su publicación en el número especial sobre Ontologías Médicas de la revista 'Computers in Biology & Medicine', Available online 6 September 2005.
[SHE1990]	Sheth, A. P., Larson, J. A: "Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases", ACM Computing Surveys, 22(3): pp. 183-236. 1990
[SIL2001]	Silvescu, A., Reinoso-Castillo, J., Andorf, C., Honavar, V., and Dobbs, D.:"Ontology-Driven Information Extraction and Knowledge Acquisition from Heterogeneous, Distributed Biological Data Sources" In: Proceedings of the IJCAI-2001 Workshop on Knowledge Discovery from Heterogeneous, Distributed, Autonomous, Dynamic Data and Knowledge Sources (2001).
[WIE1992]	Wiederhold, G.: "Mediators in the Architecture of Future Information Systems", IEEE Computer, 25(3): pp. 38-49. 1992

3.6 Meta Data

ACGT focuses not only on semantic integration of data but also on the discovery, integration and management of sharable information 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. This section aims at the consolidation of metadata requirements related to KDD within the scope of ACGT goals.

3.1.7 The Role of Meta Data

The term metadata has multiple meanings, depending on the point of view and field of application. Starting from the generic definition 'data about data' it has been used to mean different things by different authors coming from different application domains. In the scope of ACGT Knowledge Discovery on Databases, we can distinguish between two widely accepted usages of the metadata concept:

- Metadata for *extending* with additional information (also referred as *annotations*) primary targeted data to be processed by data-mining techniques. This is the case, for instance, of gene-expression data represented as a matrix with extra columns and rows for annotating related sources of information such as genomic, proteomic, or medical data. This is the use of metadata made by Bioconductor which contains packages specifically for metadata definition.
- Metadata for data and tools integration into the grid environment providing:
    - Discovery and access to stored data through *semantic* queries
    - Discovery and definition of data-mining services (a.k.a. operators) through a service *registry* or catalogue.

  This use of metadata falls into the domain of semantic web and grid architectures.

3.1.8 Metadata in Bioconductor

Bioconductor [GEN2004] is an open source and open development software project for the analysis and comprehension of genomic data. It is primarily based on the R programming language and 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 oligonucleotide microarrays (such as Affymetrix) and two-channel microarray data; identifying differentially expressed genes; graph theoretical analyses; plotting genomic data. 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, re-sampling, survival analysis, and time-series analysis.

In addition to the main software there is a large number of metadata packages available. They are mainly, but not solely, oriented towards different types of microarrays. They relate manufactured chip components to biological metadata concerning sequence, gene functionality, pathways, and physical and administrative information about genes. They provide high-performance retrieval of metadata based on probe nomenclature, or retrieval of groups of probe names based on metadata specifications. Both types of information (metadata and probe name sets) can be used effectively to extract the expression values for the named probes and metadata. Metadata packages are also provided for associating microarray and other genomic data coming from web databases such as GenBank, LocusLink and PubMed.

Bioconductor is designed as a system of interacting modules. Modularization occurs at the level of data structure, R function and R package. Data structures, including metadata, are designed to possess minimally sufficient content to be useful while maximizing efficient programming. Functions are written to do one meaningful task and no more. Packages dependencies and versioning are resolved by the automated distribution system. Metadata is placed into R packages which are distributed in the

same way as analytical software packages. For microarray analyses all data packages should have the same information (chromosomal location, gene ontology categories, etc), so that users can switch from one type of chip to another. As a result we only need a single set of tools for manipulating the metadata. This way of distributing data and metadata has the advantage of supporting reproducibility of analyses, as they can be characterized by the package versions used. This is not the case of online sources of metadata which are not version controlled. This is something important to take into account, as metadata tend to evolve as new knowledge is discovered and integrated into information systems and databases.

Two kinds of metadata can currently be distinguished in the Bioconductor project [REIM2006]:

- Metadata about microarray experiments. They encompass the chip manufacturer annotations, the MIAME data model for describing the experiment protocol, and phenotype information to describe sample-level metadata.
- General biological metadata. They encompass functional annotations such as GO, KEGG or cMAP. Biological metadata are mainly used for:
    - *Filtering* high-throughput data structures so that analyses are applied to a subset of the expression data (e.g. probes related to genes annotated with specific functions or participating in a given pathway).
    - *Labelling* probes, samples and experiments with extra information needed by analysis goals (e.g. classifier functions).
    - *Visualization* of assay data in a biological context, such as chromosomal location or pathway distribution.

3.1.9 Metadata in Semantic Architectures

Metadata for Workflow and Service Discovery

During recent years, a considerable effort has been devoted to workflows research and its successful application to areas such as Bioinformatics and Computational Biology (most notably the Biomoby, Bioweb and myGrid projects). The use of workflows is been increasingly adopted by researchers and organizations which, seems logical to think, will lead to the exponential creation of new workflows. As a consequence, a need for careful management and publishing methods is emerging. In particular, an adequate and convenient way of *annotating* workflows with metadata is of paramount importance for their posterior discovery among repositories of useful workflows. Due to great diversity of criteria for classifying and annotating them, the use of a common ontology seems reasonable. In this way, we could search workflows, not just based on their inputs and outputs, but on their purpose, internals and field of application characteristics.

The same reasoning applies to services published online (typically as Web Services). They need to be annotated and registered appropriately in order to allow their discovery by searching facilities. Therefore it seems reasonable to use the same mechanism for registering metadata about workflows and services, and for searching them.

Metadata for Workflow Composition

There are several metadata that can assist and validate the process of workflow composition. The first metadata we need is the one about the type of input and output data. We need it for checking compatibility between services that can be connected in a workflow, i.e. the output of one can be used for the input of the other. So we need a way to ask the system, given a service A what services are available that can be connected to its output. This would only return exact matches to A's output type. But even more useful would be to get a list of services that accept as input data types matching a *semantic* criteria. E.g. we can ask to get a list of services that take as an input DNA sequences whatever their format is. The next step would be to find converters between data types that allow connecting both initially incompatible services. In addition, metadata about can be used to automatically check that the workflow service is working properly. A way of doing this is by comparing expected output data, given an input sample data, with the actual output of the workflow execution. This testing data could be stored as metadata associated with the workflow.

Services Metadata Initial Proposal

Taking into account the previous considerations, we could outline now the foreseen elements needed for the description of services in the grid environment registry.

Service definition

- **Name**: service name
- **Provider domain**: relates this service with the individual, community or entity that provides it.
- **Application domain**: controlled vocabulary or ontological term which relates this service with an application area
- **Author**: service author. This could be either an individual or an entity
- **Description**: description of service
- **Input**:
  - List of Variables: list of input variables
- **Output**:
  - List of Variables: list of output variables
- **Parameters**:
  - List of Variables: list of function parameters.
- **Visualizations**:
  - List of visualization names: list of output data visualization methods
- **Admin data**: information needed for administration and management of services

Variable definition:

- **DataType**
- **Name**
- **Description** (optional)

Visualization definition:

- **Name**
- **Description** (optional)
- **Input**

- List of DataTypes

DataType:

- **Name**: type name
- **Primitive**: either true (if this is a primitive data type) or false (if this is a composite data type)
- **Attributes**:
  - List of Variables: list of attributes composing a non-primitive data type
- **Description**: data type description and/or purpose
- List of Visualizations

Admin data definition:

- Maintainer
- Operating System
- Library dependencies
- Main provider
  - Provider
- Mirrors
  - List<Provider>

Provider:

- URL
- Availability: percent of time available online
- Capacity: bandwidth, computational power

Service definition example:

```
Name: solveArrayReplicates
Provider domain: es.uma.prep
Application domain: microarray_preprocessing
Description: Solves replicates from two slides belonging to
the same experiment.
Input:
     Variable:
          Name: firstSlide
          Type: Expression matrix
     Variable:
          Name: second slide
          Type: Expression matrix
Output:
     Variable:
          Name: replicatesSolvedMatrix
          Type: Expression matrix
Parameters
     Variable:
          Name: gridDistribution
          Type: Enumerated (row_major, column_major);
     Variable:
```

```
            Name: width
            Type: Integer
      Variable:
            Name: dyeSwap
            Type: Boolean
Visualizations
      Name 1: GR graph
      Name 2: MA graph
```

Workflow Metadata Initial Proposal

Workflow metadata definition

- **Name**: workflow name.
- **Provider domain**: relates this service with the individual, community or entity that provides it.
- **Application domain**: controlled vocabulary or ontological term which relates this service with an application area
- **Author**: service author. This could be either an individual or an entity
- **Short Description**: brief description to be shown together with workflow name
- **Long Description**: formatted detailed description of workflow to be displayed in help or manual screens.
- **Workflow definition**: workflow definition written in the workflow enactor language such as Taverna's SCUFL.
- **Workflow language**: language used to define the workflow. E.g. SCUFL, Wf-XML.
- **Service Composition:** services that compose the workflow
- **Workflow Image**: image showing workflow diagram to facilitate visual inspection of workflow content. See Figure 15.
- **Visibility / Maturity**: determines whether the workshop is ready to be shared with the grid community or just used privately.

Workflow application metadata definition

- **User:** name of the user or entity that initiated the execution of the workflow
- **Input description:** semantic description (in terms of the ACGT master ontology) of the data that the workflow was executed on, plus a description of the statistical properties of the data set (mean, variance, missing values,…)
- **Input Sample**: sample data of the workflow's input
- **Output Sample**: sample data of the workflow's output

- **Objective Performance**: automatically collected measures of the performance values (e.g. predictive accuracy, cluster separation, regression error) that the workflow achieved on the input data (if applicable)
- **Subjective Performance**: free-text description of how statisfied the user is with the results provided by the workflow
- **Publication**: link to a published text that describes the results generated by this workflow application (optional)
- **Administrative information**: information about the execution of the workflow, e.g. host names, involved Grid resources, running times etc. (optional)

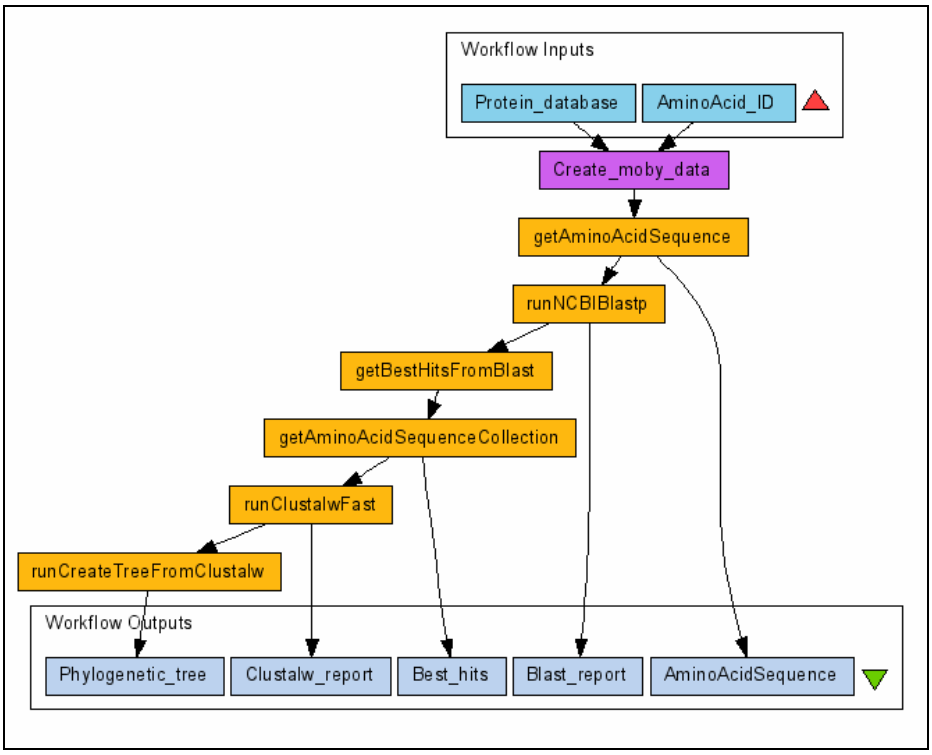


Figure 15 Workflow Image

There are currently several technological solutions available for implementing systems offering catalogues of services and workflows, such as BioMoby (WIL2002), BIOWEP and MyGRID, which already provide a wide variety of choice for accessing online analysis and data retrieval methods. Therefore, it would be an asset, for ACGT and the scientific community in general, to develop a system that is able to support different workflow languages (e.g. SCUFL, BEPL, XPDL) and different kind of services (e.g. Gridge Services, BioMoby, local commands (such as R scripts), SOAP services, WSRF) that would allow maximum flexibility for incorporating existing and new services and workflows into the ACGT grid environment. For this purpose, we need metadata that defines the type of workflow language and the type of service.

Metadata about input and output sample data is useful for automatically checking that the workflow service is properly working. This checking would be scheduled to be performed periodically (e.g. every day or every week) by the workflow manager.

3.1.10 References

[REIM2006]	Reimers, M., Carey, V., "Bioconductor: An Open Source Framework for Bioinformatics and Computational Bi8ology", Methods in Enzimology, Vol.411, 2006
[GEN2004]	Gentleman, R., et al, "Bioconductor: open software development for computational biology and bioinformatics", Genome Biology 2004, 5:R80
[NAV2005]	Navas-Delgado, I., et al, "Intelligent client for integrating bioinformatics services", Bioinformatics Vol. 22, n. 1, 2005
[WIL2002]	Wilkinson, M., "BioMOBY: An open source biological web services proposal", Brief Bioinform 3: 331-341, 2002
[ROM2006]	Romano P, Bartocci E, Bertolini G, De Paoli F, Marra D, Mauri G, Merelli E, Milanesi L, "Biowep: a workflow enactment portal for bioinformatics applications", BMC Bioinformatics, online access
[HULL2006]	Hull D., et al, "Taverna: a tool for building and running workflows of services", Nucleic Acids Research 2006 34(Web Server issue):W729-W732; doi:10.1093/nar/gkl320 online access

3.7 Managing and Sharing Knowledge

Data mining methodology and technology has been developed to date for classical business, finance, and customer-oriented application domains. Such domains are characterized by the availability of large quantities of data in an attribute-value based representation, high ratio of examples over attributes in the data sets, and weak background knowledge about the underlying entities and processes. For biomedical data these conditions do not hold. With high-throughput technologies like microarrays (for gene expression profiling) or mass-spectrometry (for proteomic profiling), a deluge of respective genomic/proteomic data is produced - about 10^4 gene, or mass/charge (m/Z) features. In addition, legal, ethical and practical limitations in clinical trials make it cumbersome to acquire a high number of patients in a clinical trial - a typical (preclinical) cohort, for research purposes, may contain only about 100-200 patient cases. Under these conditions, standard statistical and machine learning methods are likely to over-fit the structures in the data, such that a high amount of domain knowledge is needed to guide the analysis and guarantee the validity of the extracted knowledge.

It follows that the challenges of knowledge discovery in bio-medical data differs significantly from the original problems of data analysis that prompted the development of Grid technologies, for example in particle physics and astronomy: instead of the analysis of huge data sets, with the primary problems of distributed data storage and access, and the parallelization of single analysis steps, the problem here lies in the analysis of many small data sets with a plethora of possible analysis workflows. The central factor here is to make effective use of the distributed knowledge of the involved research communities in order to compensate the low

statistical significance which results from small sample sizes and complex hypotheses. Valuable kinds of knowledge include:

- Knowledge about the semantics of the data: it is well known that in data mining finding an optimal representation of the data is central for obtaining good results. That is, great care must be taken in the step of feature selection and construction [LIU1998]. An ontology-based description of the data can, for example, be used to select promising genes based on prior knowledge or construct features representing groups of similar genes. As a full feature selection is usually unfeasible for all but trivial data sets, this can greatly improve the search for good data mining solutions.
- Knowledge about the plausibility of results: when there is not enough statistical information about the validity of a hypothesis, one can look for external evidence for or against this hypothesis, because usually much more knowledge about a certain entity – gene, protein, organism, illness – is available than what is encoded in the specific data set. However, in general this knowledge is only available in the form of scientific publications – journal articles, textbooks, technical reports, memos - such that it cannot be readily integrated into an analysis workflow. Text mining technologies can alleviate this problem by structuring large collections of documents and showing up interesting connections. For example, if an analysis has identified a certain gene to be responsible for some clinical effect, one might be very interested in what is known about this gene and if there are any other publications that support or contradict this finding.
- Knowledge about analysis workflows, in particular about which workflow is optimal for a given problem. This problem is an instance of the field of workflow mining, which is data mining applied to workflows. While workflow mining usually is concerned with the reconstruction of workflows from event logs [AAL2003], for knowledge discovery we are more interested in finding a workflow with maximal predictive performance. Unfortunately, the general problem of mapping data sets to an optimal learning algorithm (known as meta learning) cannot be solved satisfactorily, both for theoretical limitations and for the great complexity problem. Hence, the Mining Mart project [MOR2004] has proposed a instance-based solution, where best-practice solutions to typical data mining problems are stored in a public database and the system supports the easy adapatation of the generic workflows to specific solutions. The integration with Grid technologies suggests an ideal combination, where the computing power of the Grid is utilized to iteratively optimize and adapt the best known solutions to a given problem.

Therefore, the main challenge for the knowledge discovery side of the ACGT project is the sharing of knowledge, either in the form of the integration of existing knowledge to design and select appropriate analysis tools, or to manage the discovered knowledge in order to make it available to the research communities. The efficient management of the different views and expertise of clinicians, biologists and data miners is of crucial importance.

In order to facilitate the sharing and re-use of workflow-related knowledge, a repository for the storage and querying of meta data, workflows, and models must be implemented. This repository should support standard languages for data mining, see e.g. [CHA2004,GRO2002,RAS2004], and should be integrated into the ACGT architecture and be able to store and query information about workflows, their

executions and their results (see also the knowledge management scenario in Chapter 5).

The workflow registration process should be supported by a user friendly interface. Web forms have been proved effective implementing such an interface as they can provide secure and ubiquitous access to workflow registries.

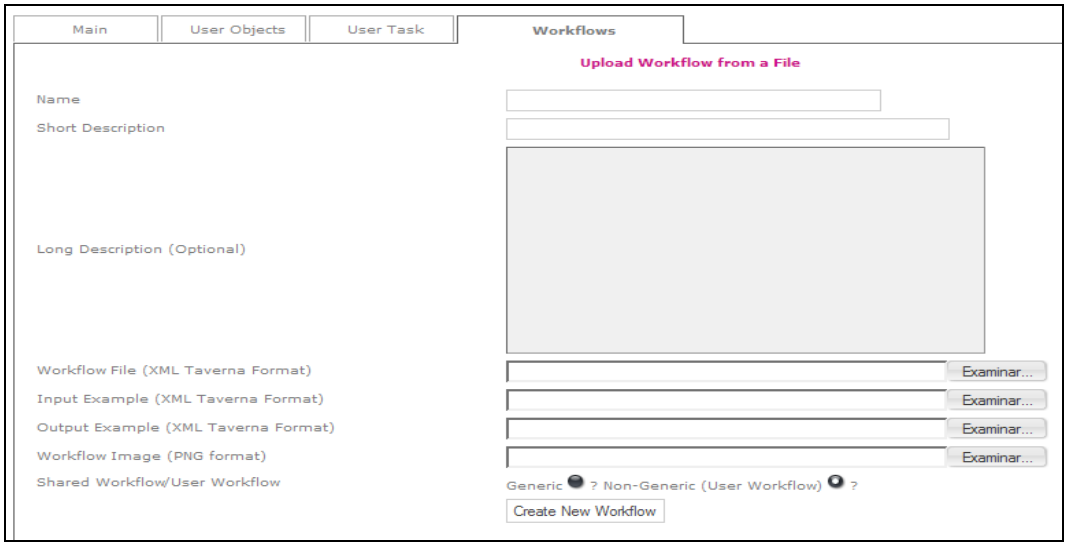


Figure 16 Web form user interface for workflow registration. This is a MOWserv web form used to register workflows in the Spanish Institute of Bioinformatics (INB).

3.1.11 References

[AAL2003]	W. van der Aalst et al., "Workflow Mining: a Survey of Issues and Approaches", in: Data and Knowledge Engineering, 47 (2), pp .237-267, 2003.
[CHA1999]	Chapman, Pete, Clinton, Julian, Khabaza, Thomas, Reinartz, Thomas, and Wirth, Rüdiger (1999), "The CRISP-DM Process Model".
[GRO2002]	Grossmann, R. Hornick, M., and Meyer, G. (2002), "Data Mining Standards Initiatives", *Communications of the ACM*, 45(8).
[LIU1998]	H. Liu and H. Motoda (eds.), "Feature Extraction, Construction and Selection: A Data Mining Perspective", Kluwer, 1998
[MOR2004]	K. Morik and M. Scholz, "The MiningMart Approach to Knowledge Discovery in Databases". in: Ning Zhong and Jiming Liu (eds.), Intelligent Technologies for Information Analysis, Springer, pp. 47 - 65, 2004.
[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.

4 Specific Tools and Services

3.8 Statistics Environment: R / Bioconductor

R is an open source statistical environment, which has quickly become the platform of choice for statistical research and many applied statistics project. 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.

In the ACGT analysis environment, R is used both as a user interface and as an analysis tool itself. Used as an analysis tool, the goal is to achieve a seamless integration of R functionality and the ACGT semantic data services in a Grid environment.

Bioconductor[1] is an open source and open development software project for the analysis and comprehension of genomic data, which is built on top of the R software. Roughly, it can be seen as a set of tools that allow the storage, manipulation and analysis of genomic data in R. The broad goals of the projects are to

- provide access to a wide range of powerful statistical and graphical methods for the analysis of genomic data;
- facilitate the integration of biological metadata in the analysis of experimental data: e.g. literature data from PubMed, annotation data from LocusLink;
- allow the rapid development of extensible, scalable, and interoperable software;
- promote high-quality documentation and reproducible research

Due to the diversity of the available functions in R and Bioconductor, very many possible data analysis scenarios are imaginable. In the following two very general, but at the same time very important and widely used scenarios are described.

4.1.1 Scenario 1: Support for Predictive Modeling

This scenario assumes that a predictive modeling tasks is given, which consists of

A data set in attribute-value format, where one of the attributes is marked as the target attribute Y (dependent variable) and the other attributes are the descriptive attributes X (independent variables).

An error measure L(y,y') defining the error that is incurred by labeling an observation x as y', when the true value of the target attribute is y.

[1] http://www.bioconductor.org

A set F of functions f : X -> Y, usually implicitely given by a statistical procedure or a data mining algorithm.

The predictive modeling task consists of finding a function $f \in F$ which minimizes the expected value of L(f(x),y).

The goal of the predictive modeling scenario is to automatically support the predictive modeling problem by providing a set of standard operators that are needed in predictive modeling, in particular

- *Evaluation:* Estimation of the expected error by train-test-splits, cross-validation, leave-one-out-estimation or bagging.
- *Parameter Optimization:* Automated optimization of the parameters of a data mining algorithm by testing out all or a heuristically chosen subset of all parameter values. This includes an estimation of the expected error as a sub-task.
- *Feature Selection.* Selecting a prediction-optimal subset of the descriptive attributes. This includes an estimation of the expected error as a sub-task.
- *Model Optimization:* The improvement of a model by specialized techniques like bagging, boosting, or stacking.
- *Operator Chaining:* The automatic subsequent execution of several data-mining and data preprocessing operators which form a joint analysis task.

4.1.2 Scenario 2: Comparison of Analysis Tools

This scenario assumes that a set of predictive modeling tasks (see Scenario 1) are given which are “compatible” in the sense that they have similar types features and that the same error measure is used. Additionally, two or more data mining algorithms, i.e. chains of basic transformation and data mining operators, are given, where each algorithm is suitable for solving each of the modeling tasks.

The goal of this scenario is to find whether one of the algorithms performs significantly better than another by executing each algorithm on each data set and comparing their performance using standard statistical tests. The necessary base prediction methods are available in R or other software (e.g stata, SAS, SPSS); however, they are difficult to compare and to use in a cross-validation and testing setting as their implementation, input and output format requirement and content are different. We suggest the implementation of a predictive modeling analysis environment which offers the possibility of using these methods in a standardised way; where analyses and methods can effectively be compared. This should be based on the existing R implementations when possible or de-novo implementation in R when required. This environment should allow cross-validation, bootstrapping, jacknife, leave-one-out methods to be applied to any of these approaches in a standardised way.

4.1.3 Technical Details and State of the Art

R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. For these reasons, R is quickly becoming a de-facto standard in for statistical computing. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles

and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either on-screen or on hardcopy, and
- a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

R code is either written in the R language, which makes is easy to transfer independently of the underlying platform. For computationally-intensive tasks, C, C++ and Fortran code can be linked and called at run time. R is easily extensible using a standardized package mechanism.

R can read and write data in a large variety of formats, including reading data from a database connection. As a standard format, data can be saved in an ASCII file in the R language, i.e. as a set of commands that will re-construct the data in R.

4.1.4 References

[BEC1988]	Becker, R., Chambers, J., and Wilks, A. (1988), "The New S Language", Chapman & Hall, London.
[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.
[MIT1997]	Mitchell, Tom (1997), "Machine Learning", McGraw Hill.

3.9 Data Preprocessing / Exploratoy Data Analysis

Visualization of gene expression data is extremely important for biological knowledge discovery, but the high dimensionality of both the data sets (e.g. [gene names] x [expression levels in each experiment] x [gene annotations] x [clinical information]) and the statistical information that can be inferred (behavioural trend, historical details, internal relationships with other elements and external relationship with other organisms) makes difficult to extract meaningful information such as interesting patterns, clusters or outliers.

For this reason, it is essential to create computing systems that implement multi-layered microarray data analyses and to develop visualization and exploratory tools. These tools are aimed to represent the results of these analyses in a clear and precise way which facilitates the extraction of meaningful insights. The description of

various visualization techniques for microarray gene expression data is shown in the appendix to this section.

A given visualization is not biological knowledge itself but it shows typically an aspect of the data, which might be explored in order to extract information (or to discover knowledge). For instance, we could explore clusters of genes in order to identify those with expression profiles similar to the profiles of known genes, which could be valuable information due to the small number of genes/proteins whose functional role is known. Hence, numerous clustering algorithms and matching strategies have been developed together with several visual representations.

An important factor to be taken into account when designing new tools for analysing gene expression data is its high dimensionality. Human perceptual skills are more effective in 2D or 3D displays, but these low dimensional projections are just a partial perspective of the data to be analysed. Therefore, there is a need for generating various low dimensional projections in order to approach the multi-dimensionality of data.

Exploratory tools allow navigating through data dimensions and multiple analyses in a visual manner. Although several mechanisms have been developed to assist researchers in selecting low dimensional projections, users are frequently unable to interact with it to find out what is interesting to them. Exploratory tools allow researchers to choose the property of projections they are interested in, examine them and locate interesting patterns, clusters or outliers. However, even with these improvements, the analyses of high dimensional microarray experiment data sets remain complex. Some examples of how visualization tools may be expanded to address the complexity of the data analyses and to assist the extraction of meaningful information are shown bellow.

Underlying problem	Exploratory method	Result
Selecting mechanisms are static. Users cannot specify what is interesting to them	Scatter plot ordering methods	2D projections can be ordered according to user selected criteria
Biological database identities do not give enough information to understand the biological meaning of the result	Gene oncology browser	facilitate the study of known gene functions within a cluster
Difficulty to explore a whole set of data when researchers know the approximate pattern of gene expression that are seeking	Profile search	genes with a specified temporal pattern ca be easily identified

These are just examples of how exploratory data analysis tools can assist researchers on knowledge discovery, but is still to be determined which are actually going to be implemented in the ACGT platform.

4.1.5 Technical Details

Tools to Preprocess Microarray Data

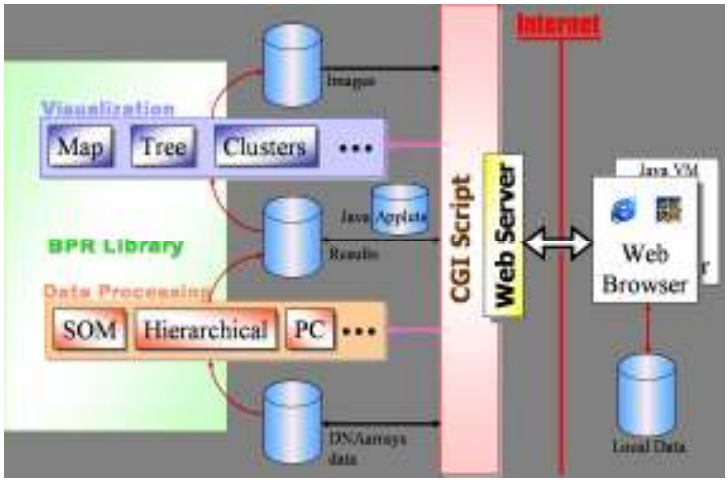
PreP is a visual tool for the pre-processing of microarray data. It encompasses a collection of methods for error detection (through visualization) and correction. The following visualization methods are used for pre-processing gene expression slides:

- **Slide view**: It is a synthetic image, built up from the available data. When data are not available, the spot is drawn with a grey mark
- **Coherent *slide* view**: It is a synthetic image, built up from the coherent data. This means that negative intensities and not available data are not shown.
- **Slide view with quality:** The quality value (supplied by replication algorithms) is shown in blue in this view.
- **AM Graph:** display the points by intensity A (axis x) and colour M (axis y). This graph allows establishing the dependency of the colour with respect the intensity, which is in fact, one of the most typical problems in normalization applications. See table X
- **AM Graph by blocks**: Display the AM graph for each of the blocks in which the slide is divided (by default using the grid). This graph allows identify variations between the different zones of the slide.
- **RG Graph**: Display the point from the red channel R (axis x) and the green channel G (axis y). The election for the channels is always red for the target and green for control.
- **Box Graph**: A box graph displays where half of the data are concentrated and makes an estimation of the possible variation, showing the data out of range independently. This graph is performed for each block in the slide, and allows visualise the differences in range for each block.
- **Values density**: Display the probability density curve of the colour value distribution. Ideally these values should be centred around cero in the normal form.
- **By block Values density**: the same as the previous graph, for each block in the slide. This allows to study the effect of the position in the slide.
- **Intensity-Intensity Graph**: Compares the intensity between two slides, giving an approximate idea about the quality of the double-scan and data dependencies between them.
- **Scatter plot of replicates**: When several points has been grouped it is possible its visualization with respect to the average of the whole set. Ideally they should be distributed around the bisection.
- **Standard deviation of the replicates**: shows the SD for each replication set with respect to the average value. Ideally they should show a linear relationship.
- **Replicates correlation**: Display the correlation between the two channels (target and control) for each replication group. Ideally it should be 1.
- **Normality of replicates**: Compute the distribution function for each of the replicate sets and drawn it with respect to the normal distribution. Ideally it should be an identity function, this is to say, it should move along the bisection.

- **Quantile-Quantile Plot**: is used to check if the dataset follows a given distribution. Currently, just the normal distribution is examined. If the data do come from a normal population, the resulting points should fall closely along a straight line.
- **Probability-Probability Plot**: similarly to QQ plots it is used to see if a given dataset follows a normal distribution by computing their p-Values.

Tools for Exploratory Analysis of Gene Expressions

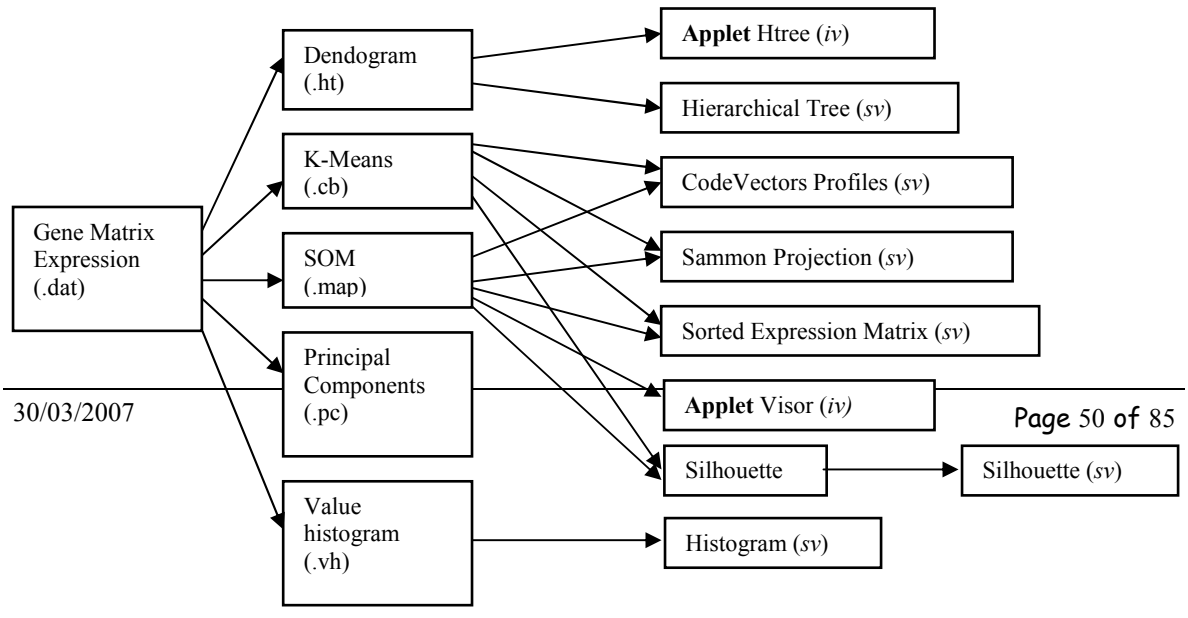
Engene is a web tool for exploratory analysis of gene expression data that aims at storing, visualizing and processing large sets of gene expression patterns. It integrates a variety of analysis and visualization tools for pre-processing, clustering and classification.



The system includes different filters and normalization methods as well as an efficient treatment of missing data. A broad range of clustering algorithms and projection methods are also provided together with an association rule generation capability. The core of Engene is a C++ library of algorithms and data handling routines wrapped by a PHP front-end. Additionally, Java applets are also available for allowing improved interaction.

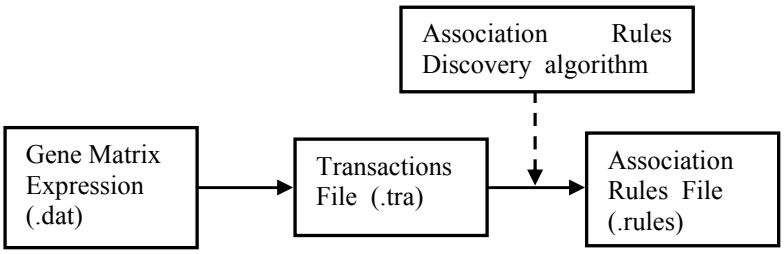
In the following diagrams it is summarized the dependencies between the different data handled and their visualization:

Gene expression and clustering visualization



sv .- Static view
iv .- Interactive view

Association rules



Exploration and Interactivity

The main means provided by Engene to allow the exploration of data analyses are (i) combination of several visualizations into one view (ii) integrated management of data analyses and files produced which allows easy navigation (iii) interactive views of data files (through the use of applets) which assist the user on inspecting and analysing the data.

Probably the most interesting part is the use of applets, which among other things allows zooming in and out, choosing colour scales and selecting and marking parts of the view and reordering of the selected parts. The HTree applet, for instance displays hierarchical clusters where you can select, invert and reorder branches in an interactive way. You can also select a given branch and display its centroid (the mean of the branch profiles), the mean distance and the standard deviation of its expression vectors. A view (colour mosaic) of the expression matrix is also available where you can select a given expression value and see the metadata associated to it.

Future work on the interaction of clustering data could involve further editing capabilities such as splitting/joining of branches/clusters and manually reassigning genes to other branches/clusters. Regarding association rules, the current visualization (which just displays a table with the rules found) could be extended to accommodate more data dimensions like expression profiles, gene ontology and annotation data.

Engene handles the following data files:

- **Data file (.dat):** a data file contains a list of vectors (data), all of the same dimension (number of variables). Moreover, a file may contain some metadata, arranged in arrays labels, variables labels and global labels.
- **Map file (.map):** contains a data (vectors) classification. This arrangement is made of outstanding vectors, the code vectors. Each vector represents a classification class. In a map file, these code vectors are interrelated by a topology. There is also additional information that associates the original data file with the classification. Each original vector might have been assigned to a code vector. To see that, for each code vector, there is a list of the indexes of the source data file original vectors. Since indexes are used, instead the vectors themselves, some operations over this file will be impossible without the original data file.
- **Hierarchical tree file (.ht):** contains a data (vectors) classification in a hierarchical binary tree. It does not contain the original data, but their references. Many of the operations on hierarchical tree files, including visualization, will need the associated data set (file.dat)
- **Principal Components Analysis (.pc):** is a quantitatively rigorous method for data reduction through the linear combination of dependent variables. All PCs are orthogonal to each other, so there is no redundant combination. This allows, for example, the projection of the original data set over a Cartesian space. The Principal Components File contains the description of the PC factors.
- **Silhouette file (.sil):** contains the silhouette value of each element. The silhouette value is a measure of the classification quality. These values lies between 1 and -1, where values near 1 represent a good classification; and values that fall under 0 are accepted as badly classified (in fact, this element is on average closer to members of some other cluster the one to which it is currently assigned. The silhouette values depend on how closed the elements of a cluster are between them and how far they are from the next closest cluster.
- **Value Histogram file (.vh):** the output of the Value Histogram Procedure is an histogram with the data distance distribution (real distances or randomise distances). This file contains such a histogram.
- **Transactions File (.tra):** a binary file containing a transaction set which is the input of the “Association rules discovery” algorithm.
- **Association Rules File (.rules):** this file is generated upon a Transactions file, by means of the Association Rules Discovery procedure (available on Engene). It contains the rules interrelating the different variables of a data file.

4.1.6 References

[GAR2003a]	García de la Nava, et al., *PreP: gene expression data pre-processing*, Bioinformatics, 2003. 19(17): p. 2328–2329
[GAR2003b]	Garcia de la Nava, et al., *Engene: the processing and exploratory analysis of gene expression data*, Bioinformatics, 2003. 19(5): p. 657-658

[PREP]	PreP site: http://chirimoyo.ac.uma.es/bitlab/services/PreP/index.htm
[ENG]	Engene site: http://chirimoyo.ac.uma.es/engenet/
[PRA2006]	T. V. Prasad & S. I. Ahson, *Visualization of microarray gene expression data*, Bioinformation 1(4), 141-145 (2006)
[SEO2003]	Seo, J., Shneiderman, B. (April 2003) *Interactive Exploration of Multidimensional Microarray Data: Scatterplot Ordering, Gene Ontology Browser, and Profile Search*, HCIL-2003-25, CS-TR-4486, UMIACS-TR-2003-55, ISR-TR-2005-68

3.10 Text Mining

Text mining as a technology addresses the need to structure, access and exploit in a systematic (computer driven) manner unstructured text, such as exists in scientific articles, patents and other documents of interest to end users.

While structuring information found in documents is a basic step, text mining [TM] main aim is to extract correlations between entities (concepts) of interests as well as trends and other secondary information that can be used in a variety of applications. A related technology to TM is Natural Language Processing (NLP) which aims to extract and organize the semantics (meaning) from natural language text. Taking an example from the life sciences, extracting from a piece of text that "*Gene A* ***up-regulates*** *Gene B*" within some context might be the output from a NLP module.

Both of these technologies represent vibrant fields of research with output including algorithms for 'understanding text, environments for developing NLP systems (e.g. GATE http://gate.ac.uk/) or actual commercial systems (e.g. Attensity's text analytics engine www.attensity.com). These technologies feed into other relatively recent fields such as *knowledge discovery* and *literature based discovery* where the goal is to identify useful information or knowledge, in the latter case using statistic, bibliometric and related techniques.

While all these technologies are very promising and significant work and results have been demonstrated in research settings, success to date is mixed in non-academic environments. Some of the reasons for this are:

- Requirement for significant 'set-up' work before the systems can perform. For example NLP systems require special grammatical, syntax and other rules to be written in order that they can process text. Often to enhance performance, special context-specific rules need to be added in order to deal with the complexities of particular fields. A very good case at hand is once again the life science field where polysemy (a single word having many meanings depending on context) often leads to complications of interpretation.
- Performance degradation as size of corpus increases: simply put, this means that most of the existing systems do not scale as corpora approach sizes such as for example are found in professional settings (e.g. the MEDLINE corpus containing 17+ million abstracts which increase at the rate of circa 5k/day).
- False sense of capabilities can lead to eventual rejection: Because the results of NLP can be impressive, yet in the eyes of the non-expert end-user

achieved in an 'unfathomable way', their error tolerance is small and so a few mistaken results often lead to (possibly unjustified) rejection of the system as a whole.

4.1.7 Scenario: Text Mining and Knowledge Discovery in a Clinical Trials Setting

From an IT systems support perspective, clinical trials represent a setting that can in theory make good use of novel technologies, since the need to enhance the clinician's ability to discover cures or treatments for conditions that existing medical knowledge cannot deal with effectively is pressing.

The complexity of this setting means of course that no single IT technology can offer an all-encompassing solution and it is with this in mind that ACGT aims to develop an environment where access to the various tools will be seamless and their use minimally disruptive to the clinicians' normal daily activities.

There are significant challenges to be met in connection with the above overall objective and from the point of view of TM and KD the following broad parameters have been agreed upon (in collaboration with the main ACGT user partners) for the initial implementation:

- The TM and KD resource will be one of a number of available elements of the clinician's toolbox for understanding disease mechanisms and researching potential cures and treatments
- The KD resource will complement 'wet' methodologies and tools as well as other s/w tools available to the clinician, the goal being to help create an overall picture of the problem at hand.
- The resource should be able to support not only the 'literature analysis' process but also the collaboration of clinicians (most probably based at remote locations) and other 'information related' activities
- The resource should be easy to use
- The resource should help generate, confirm or refute hypotheses or results obtained from other resources/tools available to the clinician
- Clinician maintains control: given the criticality of the context, the resources must be transparent to the users, offering full explanatory (drill down) capabilities that will allow the clinician to 'understand' and evaluate recommendations on the basis of access to the primary data

Within ACGT the BEA resource is available and will be dully developed and modified to meet the above challenges. This is presented in more detail in the sections that follow.

4.1.8 Technical Details

Biolab Experiment Assistant (BEA) is a literature-based environment that supports researchers in exploring problem areas of their choice, discovering hidden links and designing their experimental strategy. By integrating and cross-correlating a number of aspects of relevance to experiment design in a single environment, BEA provides multidimensional coverage of an area of interest and supports the user in designing their experimental strategy in a comprehensive manner.

The BEA environment consists of 4 elements, as shown in Figure 1: the BEA Application, BEA Documents, the BEA Reader and one or more BEA Exchange Servers.

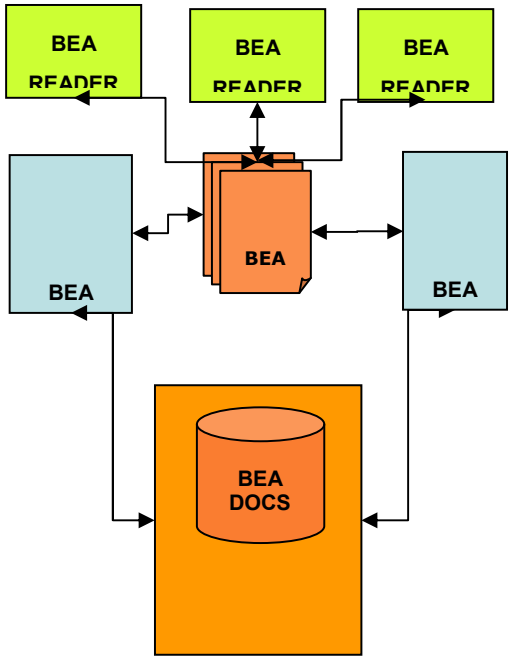


Figure 2.1: Architecture of the BEA Environment

- **BEA Application**: this is the main software module that contains the database of cross-correlated concepts (e.g. genes, pathways, experimental methods, reagents etc.) extracted from the scientific literature and allows the user to create BEA documents.
- **BEA Documents**: an exportable, interactive document that contains a variety of information capturing the work of a user during a session with BEA.
- **BEA Reader**: a stand-alone application that is a cut-down version of BEA and allows a third person to read and further annotate BEA Documents.
- **BEA Exchange Server**: a web based server containing a database of BEA documents that users can upload, search and download as they wish.

In the following we discuss the BEA tool in slightly more detail (see also BVA-ACGT-6-1v1).

BEA is a software system that supports researchers in knowledge discovery and early stage experiment strategy design. The heart of BEA is a database of concepts cross-correlated on the basis of their co-occurrence within the full text of scientific articles found in the top Science Citation Index biotech and medical journals. Currently BEA extracts and correlates concepts in over 20 classes including: genes, pathways, diseases, cell lines, organisms, experimental procedures, reagents,

medical tests and adverse events. In addition to this, the BEA database contains all patents in the health-related categories of the USPTO.

BEA is built on the premise that by making the hidden links between disparate concepts easy to identify, scientific literature analysis can help researchers understand *non-obvious interactions* and relationships that can lead to better understanding of biological mechanisms, new hypotheses and more systematic drug design and development.

Given any one or combination of the above, BEA automatically extracts the correlations from the full text literature. BEA uses text mining, NLP technologies and ontologies combined with manual curation to automatically extract the data from over 80,000 full text articles of currently 20 of the top life science journals, over 16.5 million abstracts from Medline as well as over 200.000 life science related pated from the USPTO.

Uses of BEA

BEA will return all correlations extracted from its corpus. Below are some typical questions answered by BEA, beginning with the user need to design experiments around a particular gene.

I am interested in gene X:

> What cell lines are used to study gene X? > What diseases and pathways correlate with this gene in model organism Y? > Show me the experiments and reagents people used to work with gene X > Given gene X and pathway Y, what experimental procedures have been done? > Etc, using any of the parameters listed earlier.

BEA helps users:

- explore and understand a research area of interest to them
- perform a comprehensive *browse-directed* literature review – you explore networks of associations between multiple parameters (genes, pathways, diseases, experimental methods, reagents etc.) and relevant literature is collected for you automatically
- rapidly assemble an overview of a problem area covering research, people, method/reagent and intellectual property issues.
- annotate the knowledge space they have been exploring with free text notes which they can exchange with their collaborators
- review the abstracts of selected articles and highlight areas of interest (you can collect and organize relevant quotations)
- use findings from other researchers to get a quick start in their problem area
- collaborate with their colleagues
- find related work and avoid duplication, if they are a member of a large research team

Browse-directed search

One of the unique features of BEA is the introduction of browse-directed search. Browse-directed search has been developed with the unique needs of early stage research in mind, at the same time aiming to avoid some of the problems traditional search modes face.

While everybody agrees that search engines often return a lot of irrelevant results, not many consider the difficulty of the task these programs are required to perform. When seeking information we often supply a limited number of keywords and expect the engine to:

- resolve any ambiguities in the terms we use (e.g. different word senses, synonyms etc.)
- guess the context and goal of our search
- correctly characterize the documents in its corpus, and finally
- match our incompletely specified goals with available content preferably returning the results in some prioritized list.

Given this nebulous task, it is indeed a great achievement that many search engines perform as well as they do. Yet there is another way; Figure 2 presents the search task in terms of two axes: *knowing* and *finding*.

	FIND	NOT FIND
KNOW	BOOLEAN SEARCH	BOOLEAN SEARCH WITH NEGATION
DON'T KNOW	BROWSE DIRECTED SEARCH	X

Figure 2.2: Modes of searching

The grid defines four possible modes of search for content (e.g. scientific articles):

- **M1 – I know what I want to find**: this is the most common mode where we specify one or more concepts logically linked with Boolean operators
- **M2 – I know what I don't want to find**: this is like M1 only now we also specify what results we want to exclude from what is returned
- **M3 – I don't know what I want to find**: this is what we might call browse directed search where we follow a trail of related concepts and relevant content is gathered based on the browse trail we create
- **M4 – I don't know what I don't want to find**: this mode is nonsensical although it could be considered to include content that is ignored when in M3.

Most search engines operate in modes M1 and M2 and put the entire responsibility on the user to specify as exactly as possible the search parameters. The problem with these modes is that firstly they make no use of the system's knowledge of the search space and secondly they 'punish' the user who wants to find 'what else is out there that I don't know but might be relevant'.

Contrary to most existing search tools, BEA supports modes M3 and M4 and so is very well suited to knowledge discovery especially in domains such as the life sciences where researchers are under pressure to design cures for diseases when the underlying biological mechanisms are not well understood.

4.1.9 Input data (nature and format of the data)

BEA operates on a corpus of documents and using domain specific vocabularies allows the user to review and explore correlations between the terms of the vocabularies.

Queries are generated primarily via a graph-based interface and also via menu options. An API is also available for programmatically invoking queries.

The input data are of 2 types: documents and vocabularies

Documents are using the following formats:

- Medline Abstracts: XML
- Medline full text: HTML
- USPTO patents: HTML

Other formats can be handled as required; however appropriate wrappers will have to be developed for them.

4.1.10 Output data (nature and format)

BEA output consists of results to queries that depending on the specific need are forwarded to different 'output streams' that include the visualization component, the BEA document, another BEA module or external file formats such as MS Excel and .TXT. In addition BEA uses its own CBF format in order to capture BEA user sessions.

Examples of output data include the following:

- Summary Statistics for a given concept: This is a table showing for a given concept how many concepts of specific classes it co-occurs with.

- Co-Occurrence: this is a table showing for any given pair of concepts the 'strength' of their correlation.
- Term Frequency Analysis (TFA): This table shows for a set of documents the frequency of occurrence of each concept and arranges the output by concept class.
- References: This is a list of document IDs showing for a given concept in which documents it appears.

One of the objectives of ACGT is to provide an infrastructure through which available resources (such as a modified BEA for example) will be able to interact with each other in order to support a high level task of an end user.

At the present time a complete list of elements is not available and so it is not yet possible to list interactions. This will be done at a later stage and reported in a subsequent version of this report.

BEA has been designed with large database requirements in mind. BEA's database is presently according to one metric (number of rows) the 14th largest MySQL database (over 1.2 billion rows) in the world. As such it is capable of handling very large numbers of data (see section 2.1).

The co-occurrence principle which is at the basis of BEA's approach is at its core a statistical approach and so the larger the size of the corpus, the more valid BEA's correlations are. However, from a processing point of view there is no minimum size that is required for the approach to be applicable. Any non-trivial corpus size can be processed using BEA's tools and similarly any non-trivial vocabulary (>1000 terms per category with a minimal of 2 concept categories being required) is acceptable for testing purposes.

3.11 Clustering

It is now thoroughly established that clustering is a very powerful tool in Data Analysis. This domain is so vast that it is very difficult to provide the state-of-the-art in a few pages. Even the very interesting and lengthy paper "Survey of clustering algorithms" (R. Xu and D. Wunsh, IEEE Transactions on Neural Networks, vol. 16, no. 3, May 2005), does not really give a complete picture of data classification. The important aspects such as representation of complex data and statistical methods in similarity construction and interpretation of cluster structures are not clearly dealt with.

Most often, but not always, the data are given by a description of a set of objects, say O, (a training set) by a set of attributes (also called variables or features), say A. The data table crossing O by A contains a set of feature values for each object. We have to distinguish between two families of clustering algorithms: non hierarchical and hierarchical. The former produces a partition over the set to be clustered whereas the latter outputs a totally ordered sequence of partitions, each partition being deduced from the preceding one by agglomerating clusters. Mainly, clustering methods have addressed the problem of clustering a collection of objects described by numerical or Boolean attributes, and, the most frequent conceptual data representation used is of geometrical nature. In the literature on classical clustering, the following issues are but partially addressed:

- Mathematical representation of complex data (e.g. genetic sequences, structured attributes, …);
- Statistical interpretation and summary of algorithmic results;
- Comparison of several classifications on the same set of objects or attributes;
- Association between clusters of objects and those of attributes.

The solutions to the above problems are either already integrated or can be obtained in a unified way in the framework of a very rich method based on the ascendant hierarchical classification paradigm, which we propose. This approach has been working since more than thirty years and has generated many PhD theses, articles and books. Moreover, research applications in many fields such as medical sciences (genetics, biology, clinical trials), image processing (quantization, segmentation) and social sciences (sociology, psychology, education) –just to name a few- have been performed. A software called CHAVLH ("Classification Hiérarchique par Analyse de la Vraisemblance des Liens en cas de variables Hétérogènes») has been produced in collaboration with the Ecole Polytecnique de l'Université de Nantes (France). Substantial work concerning integration of CHAVLH into R environment has been done.

4.1.11 Scenario

Let us distinguish three categories of data: clinical trial data, genomic data (sequences, gene expression profiles) and diagnosis. A first objective of the analyst would be that of evaluating the contribution of each type of the data in identifying the nature of the disease and the degree of association between them. For example, in an experience relative to the liver complaints, we could recognize the nature of the disease by clustering the patients with respect to clinical measurements, and then searching for the significant association between the patient clusters and the diagnostic data. The classification of clinical attributes enables to delimitate the syndromes clearly, and the relative role of the clusters of clinical attributes with respect to each cluster of patients may be ascertained.

Interpretation of clusters obtained from genomic data is generally more difficult, as we could observe in a recent experience with haemochromatosis (a disorder affecting iron metabolism). It is interesting to link this clustering to the previous one i.e. clustering from clinical and diagnostic data. We could use a software called v-class which is devoted to this kind of correspondence and computes original discrimination coefficients.

4.1.12 Technical Details and State of the Art

The above mentioned methodology of hierarchical classification is based on a probabilistic notion of similarity between combinatorial structures. The extreme generality of this notion and the manner in which it is built enable to take into account *a priori* knowledge for comparing the elements of the set to be organized. This approach, which refers to combinatorial data analysis, employs and develops the algorithmic aspects of hierarchical ascendant construction of a classification tree, by successive agglomerations. However, this approach has not only algorithmic aspects. Its elaboration is at the intersection of three fields: 'combinatorics', 'logic' and 'non-parametric statistics'. In fact, it gives a very general view of the 'data' and of their automatic synthesis. In this respect, the descriptive attributes (or variables) are

interpreted in terms of relations on the described set. On the other hand, set theoretical and combinatorial representation is adopted for the defined relations.

Additionally, this method produces an original notion of 'statistics' for measuring statistical relationships and proximities, namely, the 'likelihood' concept. Thus, we set up the 'likelihood' notion as a part of the 'resemblance' notion and we give a probabilistic interpretation of similarity, mentioned above. More precisely, association coefficients between descriptive variables (respectively similarity indices between described elementary objects or concepts) will refer to a probability scale for the evaluation of the involved links. This principle also underlies the 'information theory' formalism, in which the higher the amount of information quantity, the more unlikely the event concerned. In our case, the events correspond to the observed relations.

A distinctive and important point of the proposed method consists of detection of 'significant' nodes and levels of the classification tree. Intuitively speaking, a 'significant' node corresponds to achievement of a class recovering a concept, at a given degree of synthesis, while a significant level determines a partition corresponding to an equilibrium state within the clustering synthesis provided by hierarchical classification.

'Classical' view of ascendant hierarchical classification (AHC) and justification of the LLA method

AHC denotes 'hierarchical classification' methods in which the classification is obtained according to an algorithm of ascendant construction by successive agglomerations. Usually the context of data representation in which such an algorithm is expressed can be set up by means of the following triplet:

$$\left(O,\mu_O,d\right) \qquad (1)$$

where *O* is a finite set of elementary objects; the positive measure μ_O assigns a weight μ_x to each element *x* of *O* and d denotes a dissimilarity or distance index defined on *O*. Most often, in the classical view, one seeks, in a more or less justified way, to represent the couple $\left(O,\mu_O\right)$ by a cloud of points in a geometrical space. On the other hand, it is important to provide the representation space with a metric in order to evaluate faithfully the resemblance between objects. Thus, *d* may be a distance deduced from a metric. More generally, by assuming the symmetry of the dissimilarity index *d* on $(O\times O)$, and a zero value of $d\left(x,x\right)$ for each x belonging to *O*, we may deduce the following table from (1):

$$\left\{d\left(x,y\right),\mu_x,\mu_y \,/\, \{x,y\}\in P_2\left(O\right)\right\} \qquad (2)$$

where $P_2\left(O\right)$ is the set of unordered distinct object pairs. The characteristic of AHC consists of extending the notion of distance (or dissimilarity) *d*, between elements of *O* to a notion of distance (or dissimilarity) δ between subsets of *O*. Thus to the triplet (1) we associate the following one:

$$\left(P,\mu_P,\delta\right) \qquad (3)$$

where *P* is the set of all subsets of *O*, μ_P is the positive measure on *P*, deduced from μ_O. Let $\square_{\square}$ denote the set of real positive numbers. The distance or dissimilarity δ can be expressed as the mapping below:

$$\delta : \left(P \times P, \mu_P\right) \rightarrow \Re_+ \tag{4}$$

Obviously, it is of importance to induce δ from *d* in a coherent manner, but there does not exist a unique construction for this crucial induction. However, formally, we always have:

$$\left[\forall (x,y) \in P \times P\right], \delta(x,y) = f\left[\left\{d(xy) / (x,y) \in \left(X \cup Y\right) \times \left(X \cup Y\right)\right\}, \mu_{X \cup Y}\right] \tag{5}$$

where the function *f* is to be defined on the set of mutual distances between weighted elements of $Z = X \cup Y$. On the other hand $\mu_{X \cup Y}$ denotes the restriction of μ_O on $X \cup Y$:

$$\left(\forall Z \in P\right), \mu_Z = \left\{\mu_x / x \in Z\right\} \tag{6}$$

The algorithm of AHC using δ to build a classification tree on *O* is a 'trivial' mathematical principle: at each step, join the class pairs that realize the minimum value of δ. However, this mathematical 'triviality' does not entail computational and statistical 'trivialities'.

AHC can be a very general and powerful tool for data analysis. Let us specify the general structure of a data table τ. The two fundamental types of structure are: (1) $(O \times A)$; and (2) $(C \times A)$. The set of rows of τ is labeled by *O* in the first case and by *C* in the second, whereas the set of columns of τ is labeled by *A* in both cases. *O* represents the set of indivisible elementary objects (or 'individuals') while *C* defines a set of classes (or 'concepts'). In both cases, *A* is a set of descriptive attributes (or 'variables'). Let *n* and *p* denote the number of rows and columns of the data table, respectively. The (i,j)th cell which is at the intersection of ith row and jth column, contains the value of jth attribute a^j for the ith individual o_i or concept c_i respectively. This 'value' may also correspond to a modal logical formula on the set of 'categories' (or 'modalities') underlying the measure scale of the variable. In our approach, the statistical-logical interpretation of the value is considered. The classical AHC approach presented above does not address the two fundamental issues: (1) *A* is a set of qualitative variables whose categories or modalities are structured by domain knowledge (given by experts); (2) classification of the set of attributes into 'significant' classes and subclasses. Such decomposition provides a very interesting alternative to factorial analysis. The set *E* to be classified may be either *A* or *O* (respectively *C*). The significance of automatic synthesis depends on two crucial aspects: 1. the relevant coding of information, and 2. the appropriate notion of proximity on the set *P(E)* of all subsets of of the set *E* to be classified.

The general proceeding of of the LLA method can be stated as follows:

DO: *E=A*

Tree(A)=LLA(E)

DO: *E=O* (respectively *C*) according to the nature of the data table

Tree(O) = LLA(E)

A very important step consists of the class 'explanation' by means of association coefficients between the subclasses of *A* and those of *O* (respectively, *C*). This can be achieved with LLA approach.

4.1.13 Input data (nature and format)

The input to CHAVLH program, (written in FORTRAN77), consists of three files in text format: (1) a data file *chavl.don*, (2) a dictionary file *chavl.dic*, and (3) a parameter file *chavl.par*. Each observation or row of the data table is read in a fixed number of records, containing a set of observed (or coded) numerical values of the descriptive variables, in the file *chavl.don*. See the following paragraph for the description of these values. The dictionary file *chavl.dic* contains the names or identifiers of the elements to be clustered – objects or variables. Each record is a character string representing an identifier. The parameters and different options of analysis are specified in the file *chavl.par*. Examples of parameters are: title of the study, choice of the set to be clustered (objects/variables), number of objects, number of attributes, type of attributes, the value of ε (see below) etc..

CHAVLH can be used to classify the set *O* of objects (or individuals) with respect to the set *A* of attributes, or, on the other hand, to classify the set *A* with respect to the set *O*. In any case, the rectangular data table - whose rows are labelled by objects and columns are labelled by attributes – has a unique format. The descriptive variables of five different types are considered in CHAVLH:

1. quantitative or numerical variables
2. logical or Boolean variables
3. qualitative nominal or categorical variables
4. qualitative ordinal variables (in this case, the set of categories is ordered)
5. qualitative pre-ordinance ("préordonnance") variables (in this case, a weighted similarity is defined on the set of categories; it is generally given by the expert)

Whatever the mathematical nature of the descriptive attributes, and whatever the structure of the data table τ (cf. above), each entry in the table consists of a real or integer number. A real number corresponds to the value of a numerical variable, and the meaning of a particular integer number depends on the nature of the column where it appears. If the column is labelled by a qualitative variable, the integer number is a code representing a modality of the variable.

A specific case of data included in CHAVLH is concerned with the horizontal juxtaposition of contingency tables. This occurs when a large category set *I* of a qualitative attribute is distributed through different category sets $J_1, J_2, \ldots\ldots, J_k$, associated with different qualitative attributes. In this situation, the rows of the data table are labelled by *I* and the columns by $J=J_1 \cup J_2 \cup \ldots \cup J_k$. The cell (i,j) contains the number of occurrences (i.e. frequency) of the categories *i* and *j*. These characteristics are specified in the parameter file *chavl.par*. The aggregation criterion of LLA is parametrized by a real number ε ($0 \le \varepsilon \le 1$), and its value is specified in the file *chavl.par* . Two particular values of ε are found to be practically interesting, namely, $\varepsilon=0.5$ and $\varepsilon=1.0$.

Clustering the set *O* of objects

Two cases are considered for clustering the objects: (1) all descriptive variables (in the set *A*) are of the same type and (2) the descriptive variables are of different types. Each of these cases is considered separately, for the sake of clarity.

Clustering the set of variables

In CHAVLH four options are considered for clustering the variables of the same type. Each option corresponds to one of the four types of variables: quantitative, Boolean, nominal and ordinal. A separate software AVARE ("Association entre Variables RElationnelles") has been built for clustering all types of variables including the pre-ordinance variables. Actually, in this software, all the variables are mathematically coded as pre-ordinance variables. There exists a software CHAVARE, which is a version of CHAVL, integrating AVARE.

4.1.14 Output data

CHAVLH outputs the results in several files:

The file *chavl.lis* contains:

- Summary of parameters;
- Polish representation of the classification tree (cf. below for the definition);
- List of the clustered elements, sorted in descending order of the "dispersion index" which measures the relative neutral character of the element with respect to the classification;
- List of values of the "local statistic" and the "global statistic" for each level of the tree.

One or several of the files *chavl.ar1, chavl.ar2, …., chavl.ar9* are output depending on the size. They contain a reduced form of the classification tree. This diagram is obtained from the semi graphical representation of a tree coded by the Polish representation. The tree is reduced to the set of "significant nodes" as determined from the distribution of the "local statistic". Different options are available for the tree representation. Each node is labeled with an integer representing the level at which it is created. The significant nodes are highlighted by a star sign.

Several working files are also produced. In particular, the file *chavl.d01* contains the lower triangular matrix of similarity indices between the clustered elements.

The 'Polish representation' of a tree consists of a sequence of positive and negative integers. A positive integer codes an element of the clustered set while a negative integer denotes a binary aggregation. The absolute value of the latter indicates the aggregation level.

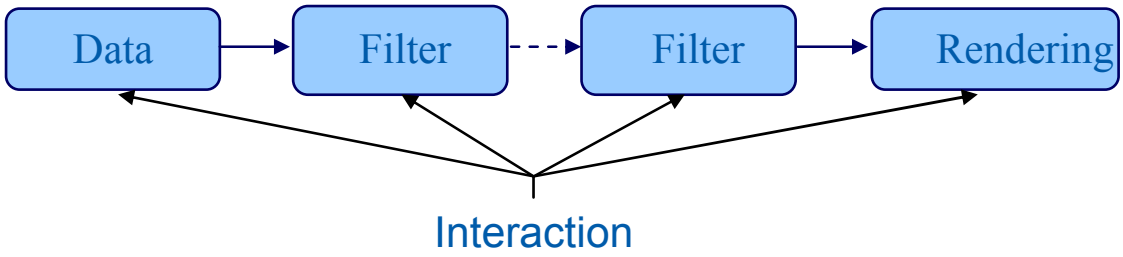
3.12 Interactive Visualisation

The discovery of novel knowledge from data is a demanding task, because there is no way to know what complex structures may be hidden in the data. Humans are very good at discovering unexpected structures by a visual inspection of the data, given that an adequate graphical representation is available. Hence, interactive visualization, which allows the user to "play" with the data until he discovery interesting patterns is a very important tool for data analysis. In addition, 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.

In cases where suitable interactive visualization methods are not readily available, a flexible framework will be available that allows for the rapid construction of new visualization methods. Moreover, the visualization methods will be sufficiently powerful to visualize high volume data sources and/or data sources of increased dimensionality and complexity. Responsive methods will be provided to interactively explore visualizations, as well as methods for collaborative visualization on geographically distributed locations.

4.1.15 Technical Details and 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.

Although the use of Grid resources for the purpose of scientific visualization has been explored in various projects [BRO2004a,CHA2004], the encapsulation of *interactive* visualization facilities into first-order Grid resources is a relatively new, but active research area [SLO2003,BRO2004b,SCA2006,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.

4.1.16 Description of the tools to be integrated into ACGT

- VTK - The Visualization Toolkit. The Visualization ToolKit (VTK[2]) 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.
- 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-linesupport for collaborative visualization across distributed sites.
    - support for collaborative visualization across distributed sites.

The activities outlined in this section will be implemented in the following phases:

[2] http://www.vtk.org/

- Implementation of the tuple space infrastructure for interactive visualization
- Integration of the "vtkfly" visualization tool into the infrastructure
- Implementation of additional visualization methods for selected scenarios
- Integration of the visualization framework into the ACGT integrated environment.

4.1.17 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.

3.13 Pathway Mining

The clinical use of microarray technologies has not yet matured to "production level" for key diseases like cancer. Even though substantial progress has been made since the inception of the technology less than ten years ago, it has still hard to convincingly beat more conventional markers. Additional progress is linked with

development of hardware, and very importantly, further development of algorithms/software tools that use array spot quality measures in processing spot and array replicates, as well as error models that reduce the importance of low-quality spots in the analysis. Also important is to develop methods for merging data from different platforms and technologies. The last issue is important since there now exist several measured cohorts under different conditions with modest sizes, which are of large potential value for calibrating diagnostic systems.

Major improvements are expected when it comes to interpreting microarray data by including prior contextual knowledge, e.g., ontologies and pathways and other annotation groups of interest. By using prior knowledge about e.g. pathways, one could infer cellular signaling pathway activity from tumor microarray data, on a sample-by-sample basis. Furthermore, one could examine whether the pathway activity of individual samples is associated with clinical classifications of the samples. This amounts to map the microarray onto a, e.g., virtual "pathway chip". By employing such encodings, one also becomes less vulnerable to noise in data. Such approaches requires two types of activities; building of prior knowledge databases, couple these to existing databases (e.g. BASE) and developing data mining tools to deal with the new encodings. This research field is relatively new. Over the last two years a few attempts on the algorithm side have been published, which all look very promising.

4.1.18 Scenario

In this scenario we assume the presence of a microarray dataset. The goal is to map the expression of a given sample or a set of samples to possible cellular signaling pathways. The output from this exercise would be a set of possible active pathways with associated statistical significance values.

One can ask for active signaling pathways given a set of “candidate genes”, given as output from previous microarray analysis step. Another possibility would be to map the expression profile of a single sample to possible active signaling pathways. This scenario requires the presence of pathway databases *e.g.*, TRANSPATH and TRANSFAC or similar.

4.1.19 Technical Details

This section briefly describes some of the requirements and details of the method found in [LIU2006], which will be as the pathway mining tool. Reference [LIU2006] is currently under review for publication and the technical details about the algorithm will therefore be brief. In short, mediating transcription factors represents signaling pathways and these mediating transcription factors are associated with putative binding sites. Gene expression signatures that correlate to these binding sites are then taken as an indication of an active signaling pathway.

The pathway-mining tool has the following requirements:

- **Databases:** Databases for signaling pathways and transcription factors is required. The method developed in [LIU2006] used TRANSPATH and TRANSFAC as such databases. The drawback is that both TRANSPATH and TRANSFAC are commercial. Other suitable free databases may exists that can be used in the ACGT environment, but the feasibility of this option is still not fully explored. UniGene and ACID [RIN2004] databases are needed to identify transcription factor binding sites.

- **Tools:** The Motifscanner (part of the Toucan software [AER2005]) software is needed for the association of gene signatures and the transcription factor binding sites.
- **Statistical analysis:** Software for doing all the statistical analysis is needed.

The implementation plan for the pathway-mining tool can logically be divided into two major steps:

- Databases for signaling pathways and other groupings of genes relevant for breast cancer containing, *e.g.*, receptors, transcription factors, and downstream targets. These databases will be integrated with BASE (except for commercial databases that may be required).
- Implementation of the pathway-mining tool developed in [LIU2006] as a plug-in to the BASE environment. A long term goal would be to provide the tool as an R function.

4.1.20 State of the Art

This research field is relatively new. A few attempts for pathway mining can be found in the work by Breslin et al. [BRE2005], Rhodes et al. [RHO2005], the EASE software [HOS2003] and Liu et al. [LIU2006].

4.1.21 References

[LIU2006]	Y. Liu and M. Ringnér. Revealing signaling pathway deregulation by using gene expression signatures and regulatory motif analysis. LU-TP 06-36 (submitted)
[BRE2005]	T. Breslin, M. Krogh, C. Peterson and C. Troein. Signal transduction pathway profiling of individual tumor samples. *BMC Bioinformatics* 2005, 6:163
[RIN2004]	M. Ringnér, S. Veerla, S. Andersson, J. Staaf and J. Häkkinen. ACID: A database for microarray clone information. *Bioinformatics* 2004, 20:2305-2306
[AER2005]	S. Aerts, P. Van Loo, G. Thijs, H. Mayer, R. de Martin, Y. Moreau and B. De Moor. TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Research* 33: W393-W396.
[RHO2005]	Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Barrette TR, Ghosh D and Chinnaiyan. AM. Mining for regulatory programs in the cancer transcriptome.*Nat Genet* 2005, 37:579-583.
[HOS2003]	Hosack DA, Dennis G Jr, Sherman BT, Lane HC and Lempicki RA. Identifying biological themes within lists of genes with EASE *Genome Biol.* 2003;4(10):R70.

3.14 Workflow Mining

Workflow mining means data mining applied to workflows. While workflow mining usually is concerned with the reconstruction of workflows from event logs [AAL2003],

for knowledge discovery we are more interested in finding a workflow with maximal predictive performance, that is the knowledge that should be extracted is about analysis workflows, in particular about which workflow is optimal for a given problem.

4.1.22 Scenario

A biomedical researcher uses the ACGT infrastructure to analyze his data. He has a semantic description of his data and of the task he wants to solve, but does not know which services and workflows are available and wants a recommendation for a workflow to solve this task (this scenario is described in more detail in Chapter 5).

4.1.23 Technical Details and State of the Art

The general problem of mapping data sets to an optimal learning algorithm or ranking learning algorithms with respect to the expected performance on a data set is known as meta learning [CHA1997,PRO2000,PFA2000]. It has for example been the focus of the European research project METAL.

Unfortunately, the problem of meta learning can in general not be solved satisfactorily, both for theoretical limitations and for the great complexity problem. On the theoretical side, the so-called No-Free-Lunch-Theorem [WOL1997,WOL1997] shows the infeasibility of the problem. On the practical side, although there is much support for massive distributed computing in current Grid environment [THA2005], it is impractical to search through even a reasonably constricted subset of the space of possible workflows for the optimal combination. Notice that the problem of constructing an optimal workflow consists not only of selecting the optimal learning algorithms, but finding a good representation and feature transformation is considered an even more important step in the mining process [CHA1999,PYL1999]. Hence, there is not only a large set of known learning algorithms to choose from (Data Mining packages like R [RDC2005], Weka [WIT2005], or Yale [RIT2001] already come with several hundred learners implemented), but also the infinite space of possible feature transformations.

In consequence, several researchers, e.g. the Mining Mart project [MOR2004] or [BAR2000], have proposed a instance-based solution. The Mining Mart project proposes to store best-practice solutions to typical data mining problems in a public database and the system supports the easy adaptation of the generic workflows to specific solutions. The integration with Grid technologies suggests an ideal combination, where best practice solutions are used as starting points and the computing power of the Grid is utilized to iteratively optimize and adapt the best known solutions to a given problem. The can be supported either by the parallelization of easily parallelizable computation tasks, such as parameter optimization, cross-validation, or feature selection, or by the speedup of single algorithms by distribution of computation, e.g. clustering [THA1999] or association rule mining [Zaki1999].

In order to support workflow mining, standard representation languages for mining tasks, workflows and datasets have to be utilized, in order to provide easily understandable and exchangeable representations of the knowledge encoded in analysis workflows. See e.g. [GRO2002,RAS2004].

4.1.24 References

[AAL2003]	W. van der Aalst et al., “Workflow Mining: a Survey of Issues and Approaches”, in: Data and Knowledge Engineering, 47 (2), pp .237-267,

	2003.
[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.*
[CHA1997]	Chan, P. and Stolfo, S. (1997), “On the accuracy of meta-learning for scalable data mining”, *Journal of Intelligent Information Systems*, 8:5-28.
[CHA1999]	Chapman, Pete, Clinton, Julian, Khabaza, Thomas, Reinartz, Thomas, and Wirth, Rüdiger (1999), “The CRISP-DM Process Model”.
[GRO2002]	Grossmann, R. Hornick, M., and Meyer, G. (2002), “Data Mining Standards Initiatives”, *Communications of the ACM*, 45(8).
[KAR1999]	Kargupta, Hillol, Huang, Weiyun, Sivakumar, Krishnamoorthy and Johnson, Erik (1999), “Distributed Clustering Using Collective Principal Component Analysis”, *Knowledge and Information Systems.*
[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.
[PFA2000]	Pfahringer, B. and Bensusan, H. and Giraud-Carrier, C. (2000), "Meta-learning by landmarking various learning algorithms", in: Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann, 743-750", 2000.
[PRO2000]	Prodromidis, A. and Chan, P. (2000), "Meta-learning in distributed data mining systems: Issues and Approaches", in: Kargupta and Chan (eds.) Advances of Distributed Data Mining, AAAI 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.
[WIT2005]	Witten, Ian and Frank, Eibe (2005), "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco.
[WOL1995]	Wolpert, D. and Macready, W. (1995), “No Free Lunch Theorems for Search”, Technical Report, Santa Fe Institute.
[WOL1997]	Wolpert, D.H. and Macready, W.G. (1997), “No Free Lunch Theorems for Optimisation”, IEEE Trans. on Evolutionary Computation, 1, pp 67-82.
[ZAK1999]	Zaki, Mohammed (1999), “Parallel and Distributed Association Mining: A Survey”, *IEEE Concurrency,* 7(4), 14-25.

5 Scenarios for Development and Validation

As was already mentioned, a number of scenarios have been proposed and described in Deliverable D2.1. These scenarios provide a guiding thread for the

development of the ACGT infrastructure by listing typical requirements and tasks that occur in clinical studies. Expectedly the ACGT clinical trials (TOP and SIOP studies) will be the basis of the first real application of the environment under development. However at the time of writing of the present document, the data of those clinical trials are not yet available, the work in WP6 will thus initially be based on selected technology-driven scenarios, while it will be extended to ACGT clinical trials as data become available.

Not all scenarios defined in D2.1 have a direct relation to the tools developed in WP6, besides some redundancy was identified between some of them. The subset of the scenarios retained is:

- Scenario SC3 (ICGKD): Based on published data, it shows the use of tools developed in relation with ACGT. This scenario focuses on the discovery of families of genes related to the phenotype of patients. The tools used involve clustering, association-rule mining (frequent itemset), and feature selection coupled with classification/prediction.
- Scenario SC1 (TOP trial based): This scenario uses a vast number of tools generally occurring in clinico-genomics data-analysis context, namely: database access, data pre-processing (e.g. gene mapping, microarray normalization), data analysis (clustering, classification, visualization), genomic annotation service, pathway analysis and text mining tools. The data needed for testing purpose will be made available by the TOP trial partners. A light version of the scenario has been proposed to orient the initial developments.
- Scenario SC2 (SIOP trial based): This scenario illustrates the use of a immunoscreening in a clinical trial context. The scenario involves uploading data in user-defined format, sequence-related analysis (sequence match against databases, query in protein databases, identification of common domains,…); immunogenics data will also be correlated with clinical data to identify genomic patterns. The data needed for testing purpose will be made available by the SIOP/Nephroblastoma trial partners.
- Scenario SC6 (Farmer, published-data based): This scenario is based on the analysis environment R, which is one of the important components of the WP6 platform. It involves uploading data to a database, then to use a number of tools built in R, namely: microarray normalization, clustering, principal component analysis, annotation-based filtering of genes and gene-set enrichment analysis.

3.15 Scenario SC3: Correlating Phenotypical and Genotypical Profiles (ICGKD)

A detailed description of this scenario is available on the BSCW/ERCIM document server for ACGT. The document describes two aspects of the scenario, one oriented towards tools development (and described elsewhere in the present document) and one oriented towards data analysis. The latter is described in greater detail here.

5.1.1 Scope of the scenario

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

The aim is: (a) to support the linkage between patients' clinical and genomic patients' profiles, and (b) to reveal interesting relations between the relevant data sources towards the composition of (potentially) interesting and indicative individualized (i.e., target-population oriented) clinico-genomic profiles.

5.1.2 Data

The realization of the scenario relates both clinical and microarray (gene expression) data.

The initial realization (feasibility study) of the scenario will be based on relevant public-domain data. In this context we will focus on the **NKI breast-cancer** study [VIJ2002,VEE2000a,VEE2000b], and on the provided dataset which records, among other, clinical, histopathology, treatment and clinical outcome as well as respective gene-expression data for 295 breast-cancer patient cases. For the relevant see link in ref. [VIJ2002].

5.1.3 ICGKD in brief

Step-1. The clinico-histopathology and the gene-expression data of patients' samples are appropriately recorded into relevant clinical (*OncoSurgury* and *HistoPathology*) and gene-expression *information systems* (i.e., the BASE system is utilised).

Step-2. A data *mediation* service is called and utilised to query and retrieve data from the respective clinical and gene-expression data sources. The results are stored into a standard-formatted *data-enriched XML* file which is indexed, stored and managed by appropriate data-warehouse operations.

Step-3 *Formation of Clinico-Histopathology Phenotypical Profiles (CHPP).* The samples are assigned to various *clinico-histopathological categories* that present and correspond to specific Clinico-Histopathological *Phenotypical Profiles – CHPP,* e.g., profiles or, *classes* referring to specific tumour types, stages, drug response statuses etc. This may be accomplished by reference to the quest specifics of a clinico-genomic research trial, i.e., targeted patient cohorts that meet pre-specified clinico-histopathology criteria.

Identification of such groups is supported by standard entries in the respective clinical information systems' patient records, and retrieved by corresponding queries. For this purpose the TNM cancer-categorization standard (TNM: 'T'umour stage, lymph-'N'ode, and 'M'etastasis status) is employed and coupled with information concerning other patients' characteristics (e.g., medical image annotations and respective entries in the HistoPathology information system) in order to form and target specific patients' phenotypes.

Clinico-Histopathology Profiles. Besides the investigator-guided process, advanced data-mining operations may be utilised in order to *automatically discover* and form *indicative CHPPs*, for example by employing and utilizing appropriate data-mining operations: (***i***) ***Clustering*** of patients/samples to categories of *similar* clinico-histopathology profiles facilitates the identification of potentially interesting cohorts

and ease the targeting of a clinico-genomic study. It is more an *exploratory* enterprise, especially in the environment of a multi-centric research clinical trial where a special need is to *"find an adequate number of patients/samples meeting specific clinico-histopathological descriptive profiles to target* (i.e., engagement of specific subjects) *and initiate appropriate statistical data analysis tasks"*. Among others, the clustering operations support *feature focusing* in order to help the investigator to focus on specific clinico-histopathology features of interest. (***ii***) In the case that the research-trials' basic question(s) is(are) formed and respective *decision-feature(s) is (are)* selected (e.g., *"metastasis vs. no-metastasis patients"*) then, with the aid of ***classification*** and ***feature-selection*** techniques and algorithms the clinical investigator may identify combinations of clinico-histopathology feature-value combinations being able to ***discriminate*** between pre-specified patient groups, e.g., *"good vs. bad prognostic clinico-histopathology profiles"*. These CHPPs could be contrasted with respective gene-expression profiles in order to investigate and assess respective diagnostic/ prognostic prediction performances (see step 4 below). Note that a special prerequisite of the employed data-mining/ knowledge-discovery operations is that, the results output they offer should be in a (semantically and syntactically) *standard format so that it could be utilised by sub-sequent knowledge-discovery operations and data-mining algorithms in a* ***workflow*** *mode.*

Step-4. *Formation of Gene-Expression Phenotypic Profiles (GEPP).* By measuring transcription (gene expression) levels of genes in an organism under various conditions, and for different tissues, we build-up *Gene Expression Phenotypic Profiles or, patterns – GEPP. The GEPPs* characterize the dynamic functioning of each gene in the genome. 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. *Corellating CHPPs with GEPPs.* The quest now is to link and corelate the two phenotypic profile types. We introduce the smooth integration of data-mining operations for this. The integration is based on a *multi-strategy* data-mining process, the ICGKD process, in the sense that it smoothly combines different data-mining methodologies, i.e, *clustering*, *association rules mining* and *feature-selection* with *classification* operations. It unfolds into the following: (***i)*** First an appropriate (*unsupervised)* ***clustering*** method is utilised in order to identify clusters of genes or, ***metagenes***, based on their gene-expression profiles. It is mainly meant to reduce the dimensionality of the search space (i.e., *from 1000^{ths} of genes to 10^{ths} of metagenes*). The metagenes present potentialy interesting GEPPs. The question then is: *"does these GEPPs relate and how with specific CHPPs?"* This task is accomplished with the aid of an ***association rules mining*** (ARM) methodology in order to *automatically discover 'highly confident' correlations between GEPPs and CHPPs.* Such a discovery may conclude to a *re-classification* of the targeted disease and reveal interesting and unexplored individualised responses and behaviours (e.e., a specific CHPP). Moreover, if the description of the identified CHPPs includes a 'decision feature' (e.g., *years-of-survival*) then, a ***gene-selection*** operation is performed just on the patients' samples that meet these CHPPs, and just on the genes in the correlated GEPPs (i.e., genes in the respective metagenes). The outcomes of this exercise are ***individualised*** molecular/genomic signatures.

ICGKD is inspired and implements a *multi-strategy machine learning* and *case-based reasoning methodology*. The whole approach, and the actual realization of the ICGKD scenario, is based on the *smooth integration* of three distinct data-mining

components: (***a***) ***Clustering***. Based on a novel *k-means clustering* algorithm operating on *categorical* data, named *discr-kmeans***.** 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 down-regulated (i.e., low) or, up-regulated (i.e., 'high') gene-expression profiles [MAY2006,KAN2006]; (***b***) ***Association Rules Mining***. It is aimed for the discovery of *'causal'* relations (rules with high confidence) between genes (actual clusters of genes, i.e., the metagenes) and patients' phenotypic profiles. The *HealthObs* system (actually its functions) is utilised for this purpose [POT2004a]; and (***c***) ***Feature Selection*** and ***classification***. For the selection of the *most discriminant genes*, i.e., genes being able to adequately discriminate between CHPP decision feature-classes on the basis of the targeted metagenes. The *MineGene* suite of microarray data-analysis system (actually its functions) is utilised for this purpose [POT2004b].

5.1.4 Services needed for the scenario

The services needed to run the present scenario are:

- Service for storage of clinical data
  - OncoSurgery
  - Histopathology
- Service for storage of gene-expression data
  - BASE
- Service for discovery of Clinico-Histopathological profile (CHPP)
  - Clustering tool (discr-kmeans)
- Service for discrimination of patients
  - Feature-selection (MineGene)
  - Classification (MineGene)
- Service for the formation of Gene-Expression Phenotypic Profile (GEPP)
  - Pattern discovery
- Service for corellating CHPPs and GEPPs
  - Clustering tool
  - Metagene construction tool (from clustering service) (MineGene)
  - Association rules mining tool (HealthObs)

References

[ACGT2006]	ACGT D2.1. (2006). *User requirements and specification of the ACGT internal clinical trial.* ACGT deliverable D2.1. https://bscw.ercim.org/bscw/bscw.cgi/d163285/ACGT_D2.1_FORTH_%20final.pdf
[POT2005]	Giorgos Potamias, 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
[GRU2003]	Gruvberger S.K., Ringner M., Eden P., Borg A., Ferno M., Peterson C., and Meltz

	Expression profiling to predict outcome in breast cancer: the influence of sample s *Cancer Res.* 5(1): 23–26.
[KAN2006]	Kanterakis A., and Potamias G. (2006). Supporting Clinico-Genomic Knowledge Discovery: A Multi-strategy Data Mining Process. In G. Antoniou, et al. (Eds.): SETN 2006, *LNAI* 3955, pp. 520-524. http://www.springerlink.com/(mibwxn45mifi5r45aa5ddp2q)/app/home/content.asp?referrer=contribution&format=2&page=1&pagecount=0
[LOP2000]	Lopez L.M., Ruiz I.F., Bueno R.M., and Ruiz G.T. (2000). Dynamic Discretisation of Continuous Values from Time Series. In R.L. Mantaras and E. Plaza (Eds.), *Proceedings of the 11th European Conference on Machine Learning*, *LNAI* 1810, 290-291.
[MAY2006]	May M., Potamias G., and Ruping S. (2006). Grid-based Knowledge Discovery in Clinico-Genomic Data. *International Symposium on Biological and Medical Data Analysis* (ISBMD 2006), 7-8 December 2006, Thessaloniki, Greece.
[PER2006]	Perreard L. et al. (2006).Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. *Breast Cancer Res.* 8(2): R23.
[POT2004a]	Potamias G., Koumakis L., and Moustakis V. (2004a). Mining XML Clinical Data: The HealthObs System. *Ingenierie des systems d'information* 10(1): 59–79.
[POT2004b]	Potamias G., Koumakis L., and Moustakis V. (2004b). Gene Selection via Discretized Gene Expression Profiles and Greedy Feature-Elimination. *LNAI* 3025, 256–266. http://www.springerlink.com/(mibwxn45mifi5r45aa5ddp2q)/app/home/content.asp?referrer=contribution&format=2&page=1&pagecount=11.
[TOP2006]	TOP-trial. (2006). The Trial of Principle (TOP Trial). http://www.clinicaltrials.gov/ct/show/NCT00162812;jsessionid=48FF824FA591095001BFFA89CF812D74?order=1
[VEE2002a]	van 't Veer L.J. et al. (2002a).Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871), 530-536.
[VEE2002b]	van 't Veer L.J. et al. (2002b). Expression profiling predicts outcome in breast cancer. *Breast Cancer Res.*, 5(1), 57-58.
[VIJ2002]	van de Vijver M.J. et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med.* 347(25), 1999–2009. [NKI study data: http://www.rii.com/publications/2002/nejm.html].

3.16 Scenario SC1: Complex Query Scenario for the TOP Trial

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).

5.1.5 Scope of the scenario

A researcher involved in the TOP trial tests a hypothesis to explain behavior of non-responder patients who were withdrawn from the trial. The detailed list of tools to be used in every step of the scenario will be given in a separate document.

5.1.6 Data

For initial testing, a subset of the patient data available in the TOP trial will be provided to WP6. Clinical as well as microarray, FISH and other molecular marker data will be available.

5.1.7 Complex Query Scenario in-brief

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 clinicogenomic/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 ultimately implies ***access to and retrieval of data*** from the respective **Clinical Information Systems** in the Cretan and JBI sites, although an *ad hoc* database may be used for testing purpose; 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.

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.

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*.

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.

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.

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

5.1.8 Services needed for the scenario

The full implementation of the TOP complex query scenario requires:

- Service for storage of sample-associated clinical data
- Service for storage of gene expression data
    - BASE or other DB (the scenario requires simultaneous access to two databases, subsets of samples must be queriable based on query on clinical data, data from different platforms should be combined)
- Service for storage of clinical data, sample meta-data
    - BASE or other
- Service for identification of annotation
    - Identification of common features on different microarray platforms; this should be the result of a workflow calling annotation services of finer granularity
- Service for biostatistics analysis
    - R (clustering)
- Service for querying genomic annotation
    - Access to public nomenclature, annotation and ontology databases
    - Service for querying public gene-expression databases
- Service for identification of molecular pathways
- Biovista, LundU tools
- Service for biomedical literature search/mining
    - BEA/Biovista
- Reporting (standard reporting form for the steps in the present analysis)

The lighter version of the scenario (TOP-light) requires:

- Service for storage of gene-expression/genotyping array data
    - BASE
- Service for storage of clinical data, sample meta-data
    - BASE or other
- Service for microarray annotation
    - Can be provided as an external tab-separated text file
- Service for data analysis: R

3.17 Scenario SC2: Identification of Nephroblastoma Antigens

The scenario aims at the characterization of the immune response against human tumors (nephroblastoma, Wilms tumors), by mean of immunoscreening of a cDNA expression library followed by the characterization of tumor-specific antigens by bioinformatics means. The scenario also addresses the possibility of using the patterns of seroreactivity as prognostic marker for chemotherapeutic response and outcome.

The scenario addresses the access to public databases and web resources from within the ACGT infrastructure.

5.1.9 Scope of the scenario

A researcher conducts an immunoscreening experiment and characterizes the outcome by bioinformatics means.

5.1.10 Data

Seroreactivity for patients and healthy blood donors will be provided by the SIOP/Nephroblastoma trial consortium.

5.1.11 Nephroblastoma Antigens scenario in-brief

Phase I: Characterization of antigens

- 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.

- 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.

Phase II: Integration with clinical data

- 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 clones are 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 tumor marker.

5.1.12 Services required for the scenario

Phase I: Characterization of Antigens

- Service for storage of SEREX data
  - Data in matrix/dataframe form, with semantic description of row/column contents
- Service for Expasy access
  - SEREX clone sequences are converted to protein sequences in the six reading frames
  - Manual selection of most relevant reading frame may/will be necessary
- Service for intermediate storage of results
  - Output of manual curation of Expasy search should be retrievable on a later session
- Service to sequence-matching/database-search (NCBI/BLAST)
  - Returns best matched protein IDs
- Service to provide NCBI protein annotation
- Service to search/access protein information in OMIM database
- Service to sequence-matching/sequence-comparison (blast2seq)
  - Output of SEREX protein sequence identification must be compared to sequences picked from databases
- Service to search homologous genes
- Service to link protein/gene IDs to relevant publications
  - Link to PubMed/BEA

Phase II: Integration with clinical data

- Service to store clinical data
  - Possibly through anonymization service
  - Individual patients will have multiple data records on different time points
- Service for statistical analysis, R
- Service for tracking individual patient evolution in antigen/antibody pattern
  - Time course

3.18 Scenario SC6: Molecular Apocrine Breast Cancer

The scenario illustrates how clinical data can be used to conduct more fundamental biomedical research. In the present case the various tools allow the identification of a subcategory of breast cancer based on microarray data.

This scenario is described in greater detail in a document available on the ERCIM/BSCW document server for ACGT.

5.1.13 Scope of the scenario

A biomedical researcher uses the ACGT infrastructure to conduct fundamental research in the context of a clinical trial. The focus of the scenario is on the use of the R/BioConductor statistical environment, including the implementation of user-defined extensions.

5.1.14 Data

The scenario uses simple, published clinical and Affymetrix microarray data. All relevant files are available on the ERCIM/BSCW document repository.

5.1.15 Molecular Apocrine Breast Cancer scenario in-brief

Data access

Microarray and clinical data are to be uploaded to an ACGT database (i.e. BASE), such that they can be retrieved later on from the inside of the database.

Quality control and normalization

Various tools available in BioConductor are called to assess the quality of the microarray data. (Alternatively other tools available in ACGT may be used.)

Data analysis

A list of steps needed to reproduce the results of the article from which the data have been taken (see ERCIM/BSCW document server) include:

- Gene filtering (using annotation tools to select a subset of features representative of a gene)
- Cluster analysis
- Principal component analysis

This list can be extended at will by using other components of the R/BioConductor environment. Alternatively, the ability to combine R/BioConductor tools with other ACGT non-R-based tools should be assessed (e.g. for clustering).

Extension of the R environment

The scenario includes a script for a method not belonging to the base R / BioConductor packages. The script conducts a Gene Set Enrichment Analysis, which assesses whether a group of (possibly related) genes is significantly more expressed in a list of genes.

5.1.16 Services needed for the scenario

The services for this scenario are essentially the same as for the TOP-light scenario.

- Service for storage of gene expression data

  - BASE
- Service for storage of clinical data, sample meta-data
  - BASE or other
- Service for microarray annotation
- Service for data analysis
  - R (clustering, PCA)

3.19 Studying the prognostic value of specific pathways for different tumours

This scenario shows how clinical data can be used to help testing the prognostic potential of pathways studied in the laboratory for different cancer sites.The profiles should also include potential therapy targets or describe a therapeutic pathway. Ideally a profile should be related to response and outcome of therapy targeting the pathway, or help to categorise patients prospectively to test the validity of the pathway. Examples would include profiles done before and after a novel therapy targeting a specific pathway eg VEGF to develop a profile to define responders. Another would be resistance to radiotherapy and relapse in the ipsilateral breast,so helping select who should have a mastectomy. A potential profile may involve hypoxia and DNA repair pathways.

In this scenario a molecular biology laboratory uses the ACGT infrastructure to understand if a number of pathways studied in the laboratory could have prognostic for survival and predictive of response to therapy in breast cancer. They need to use published and clinical data to compare signatures in terms of their biological information and clinical usefulness. This process needs to be updated as more literature and laboratory data become available.

5.1.17 Scope of the scenario

Several signatures have been suggested in previous studies which have been shown to have a certain degree of prognostic value in breast cancer. These signatures have been derived in very different ways; laboratory cell line experiments (ref), analyses of in-vivo breast cancer data (ref), or analyses of other cancer sites which resulted prognostic for breast cancer (ref). On the other end, the researchers in the laboratory are accumulating information for their pathways of interest by performing experiments on cell lines. The scope of this scenario is to mine the literature to test the prognostic potential of these pathways for breast cancer. This is a two way process: on one side published prognostic signatures are mined to find relevant pathways; on the other end signature produced in the lab are tested for their prognostic potential.

5.1.18 Data

Relevant literature should be collected and a search for cancer patient series should be performed where microarray data, relevant clinical data and treatment outcome data are all available with relevant documentation published.

5.1.19 Scenario in-brief

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

1. Data Access. (i) Use laboratory information, Biovista and pathway mining tools to build pathways specific gene lists. (ii) Get gene expression (microarray) data and clinical data for the relevant clinical studies needs to be accessed and using BASE or equivalent MIAME compliant microarray information systems.

2. Data Pre-processing

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

(ii) As very different platform will be involved; merge data where possible (e.g. different affymetrix platforms) or normalize the data separately for each study when platform integration is not possible. This implies use of gene expression normalization/transformation tools/services in R/Bioconductor and, where available, ACGT tools to integrate data across platforms.

3. Data Analysis and reporting

(i) Use pathway mining tools to understand the biology behind published prognostic signature. (ii) Use regression tools and survival analysis tools to find prognostic and clinical significance of signatures; and respective visualization tools.

4. Data update

Automatically update the results as more literature becomes available; this implies the possibility of storing, re-using (but also updating with new-methods – so some sort of flag) the workflow.

3.20 Knowledge Management Scenario

An important aspect of the ACGT knowledge discovery architecture is not only to support the creation of new analysis workflows, but also to facilitate the sharing of the knowledge that is encoded in this workflows, i.e. which approach solves which problem best. This scenario describes one approach to exploit the the information that can be gathered in an automatically generated database of workflow applications and their results.

5.1.20 Scope of the scenario

A biomedical researcher uses the ACGT infrastructure to analyze his data. He has a semantic description of his data and of the task he wants to solve, but does not know which services and workflows are available and recommended to solve this task.

5.1.21 Data

The scenario uses a database of workflow applications, where each examples of a workflow application consist of a description of the workflow, annotated with a description of the task that this workflow is supposed to solve, a description of the input data and a description of the result of the workflow (e.g. performance values in a predictive scenario).

5.1.22 Knowledge Management Scenario in brief

This scenario may be solved by supporting the execution of the following example query: Given a set of data, show me all workflows that

- have already been used on semantically similar data and
- have already been used on statistically similar data and
- solve my problem and
- have shown good results and
- are applicable to my data and
- I have the credentials to execute.

Then, rank the workflows by their expected performance and execute the 10 best workflows on my data.