Multi-Knowledge: Collaborative Environments for the Extraction of New Knowledge from Heterogenous Medical Data Sources

Michele Amoretti[1], Diego Ardigò[2], Franco Mercalli[3]

[1] Information Engineering Department, University of Parma, Italy
[2] Internal Medicine Department, University of Parma, Italy
[3] Centre for Scientific Culture "A. Volta", Como, Italy

Abstract. The general aim of the Multi-Knowledge project is to develop a collaborative environment to allow networks of co-operating medical research centres to create, exchange and manipulate new knowledge from heterogeneous data sources.
The Multi-Knowledge service-oriented architecture will enable workflow design and execution based on novel operating procedures to manage and combine heterogeneous data and make them easily available for the imputation of study algorithms. In this paper we describe the general framework and the pilot application, providing preliminary results of the combined analysis of microarray and clinical data of 50 patients.

1 Introduction

The Multi-Knowledge project [1], which is funded by the European Commissionin the context of the Sixth Framework Programme for Research and Technological Development (Project #027106, thematic area Information Society Technologies), starts from the data processing needs of a network of Medical Research Centres, in Europe and USA, partners in the project and cooperating in researches related to the link between metabolic diseases and cardiovascular risks. These needs are mostly related to the integration of three main sources of information: clinical data, patient-specific genomic and proteomic data (in particular data produced through microarray technology), and demographic data.

In this context the aim of Multi-Knowledge is the development and the validation of a knowledge management environment to allow different groups of researchers, dealing with different sources of data and technological and organisational contexts, to create, exchange and manipulate new knowledge in a seamless way. The ambition is also to create a technological and methodological frame that can easily be extended to include additional sources of data and expertise (bio-medical data, images, environmental data), and can be applied to wider sectors of medical research. This innovative scenario is illustrated in figure 1.

Critical and difficult issues addressed in the project are the management of data that are heterogeneous in nature (continuous and categorical, with different order of magnitude, different degree of precision, etc.), origin (statistical programs, manual introduction from an operator, etc.), and coming from different data environments (from the

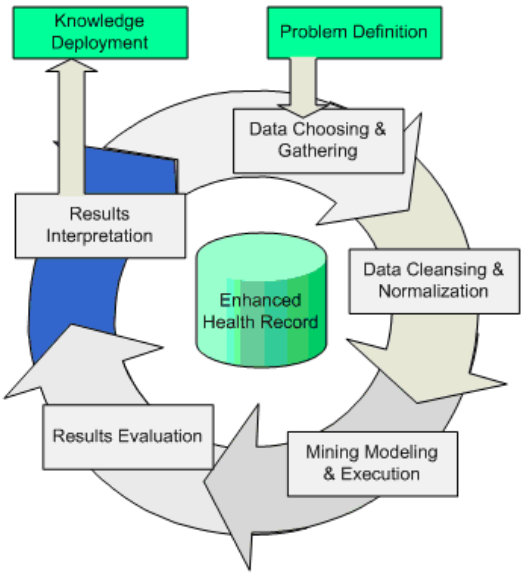

Fig. 1. Multi-Knowledge innovative scenario.

clinical setting to the molecular biology lab). The Multi-Knowledge service-oriented architecture we are developing will enable workflow design and execution based on novel operating procedures to manage and combine heterogeneous data and make them easily available for the imputation of study algorithms.

The paper is organized as follows. In section 2 we discuss related works in the field of biomedical distributed services. In section 3 we illustrate the general template for Multi-Knowledge workflows, together with some use cases. In section 4 we describe the system prototype and the pilot experiment. In section 5 we illustrate some preliminary results of data analysis experiments related to 50 clinical and microarray data samples. Finally, we outline some directions for further research and development.

2 Related Works

In the context of clinical services, the European Commission is funding two complementary projects: COCOON [2] and ARTEMIS [3]. COCOON is aimed at setting up a set of regional semantics-based healthcare information infrastructure with the goal of reducing medical errors. ARTEMIS aims to develop a semantic Web Services based interoperability framework for the healthcare domain, building upon a peer-to-peer architecture in order to facilitate the discovery of healthcare Web Services.

The Biological Web Services (BWS) page [4] describes the main services that are available as of March 2006, with appropriate links. Among the services listed by BWS, GeneCruiser [5] is a Web Service for the annotation of microarray data, developed at the Broad Institute (a research collaboration of MIT, Harvard and its affiliated Hospitals). GeneCruiser allows users to annotate their genomic data by mapping microarray

feature identifiers to gene identifiers from databases, such as UniGene, while providing links to web resources, such as the UCSC Genome Browser. It relies on a regularly updated database that retrieves and indexes the mappings between microarray probes and genomic databases. Genes are identified using the Life Sciences Identifier standard.

A more complex example of Web Service-oriented architecture providing transparent access to biomedical applications on distributed computational resources is the National Biomedical Computation Resource (NBCR) [6], which is based on Grid technologies such as Globus Toolkit. NBCR users are allowed to design and execute complex biomedical analysis pipelines or workflows of services.

Compared to these initiatives, the Multi-Knowledge project is a step forward since its objective is the creation of collaborative environments in which many kinds of actors (physicians, biomedical researchers, etc.) participate in the workflow execution.

3 Multi-Knowledge Workflows

Multi-Knowledge experiments represent process instances, and each experiment step represents an activity. Experiment steps are defined by and conducted under the responsibility of a research team, coordinated by a Principal Investigator. Starting from a patients data sample, usually defined and collected in the first work phases, the experiment is set to conduct successive data analysis cycles, aimed at extracting new knowledge through the exploitation of full integration among heterogeneous data clinical, demographical, genomic and proteomic managed by a diverse set of researchers. Biomedical researchers and biostatisticians are the major members of the research team.

The data sample is populated through the execution of relevant data collection steps. As above mentioned, data collection steps are normally the first steps to be conducted. Data analysis steps form the core of the experiments analysis cycle. Through them, the data sample is successively analysed by different classes of researchers having different ”scientific cultures” and backgrounds that use different analysis tools, work in different environments, at geographically dispersed sites. Each of the data analysis step may generate new knowledge elements that contribute to create and successively expand an *experiment-related body of knowledge (EBoK)*. Based on an analysis of the EBoK (performed from their different scientific point of views) research team members can propose the execution of additional experiment steps or to further carry on the process.

Thus, the Multi-Knowledge workflow system:

- introduces experiment steps that are conducted by different researchers with diverse scientific background and cultures;
- supports the need of passing control back and forth from different researchers to perform data analysis steps relating to completely different mathematical foundations;
- allows the experiments consist of dynamical cycles of data collection and analysis that aim at progressively achieving the scientific goal initially stated for the experiment;
- concerns the collaboration among different teams, which are independently performing experiments in related areas.

The general template for Multi-Knowledge workflows complies with the activity diagram in figure 2.

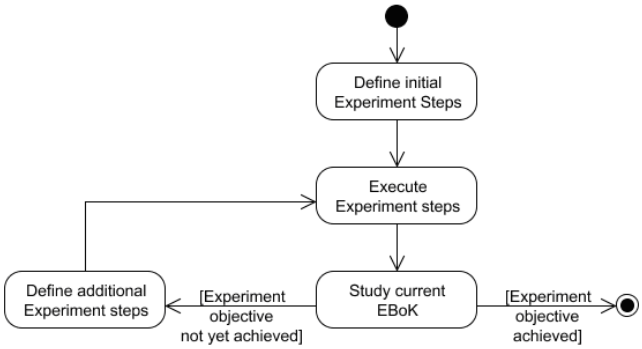

Fig. 2. Multi-Knowledge workflow template.

When a team member, possibly after receiving a suggestion sent by another team member or by the principal investigator, decides to execute an experiment step, he/she:

- revises the proposed experiment step definition and possibly improves it based on her/his specific knowledge;
- executes the experiment step;
- adds an annotation, presenting the motivations for the experiment step as well as comments on steps execution and outcome.

The workflow engine reacts by logging the experiment step that has been executed, in terms of task identifier, ask parameters and used data set, and by recording the annotations produced by the team member that executed the Step.

Moreover, team members are allowed to browse the current experiment status, consisting of all the experiment steps conducted so far and related information. The system offers a comprehensive view of this knowledge, allowing to choose and extract graphic representations of different (including intermediate) steps of the experiment, to define and print reports based on the experiment status or on specific parts of it, and to perform statistics on logging information coming from different experiments managed within the system, in order to extract performance measures and identify best practices.

4 Pilot Experiment and System Prototypes

In the first instance of MK pilot study clinical, laboratory, instrumental and genomic information has been collected from 50 subjects by the Department of Internal Medicine of the University of Parma. The sample has been used to validate the first MK IT system

prototype, in particular the system modules related to data collection and normalization, and to define preliminary OLAP/mining models for heterogenous data analysis.

On the server side, the following modules were deployed:

- the **Portal**, which is the point of access to the knowledge extraction system;
- the **Data Collection and Normalization (MK-DCNS)**, whose goal is to provide a common integration service bus and a set of specialized application adapters for the collection, integration and normalization of heterogeneous data from heterogeneous data sources.

Different kinds of biological data can be inserted into the MK-DCNS. Referring to RNA expression arrays and protein arrays, *microarray measurements* are given as a set of feature extraction (FE) files and an indication of which columns from them to use. Each FE file represents one experiment and contains all the data derived from that microarray. Each expression FE file contains data on about 40000 genes and each protein FE file contains data on about 100 proteins. Furthermore, *metabolomics data* are given as tab delimited text files with two columns. The first column contains the metabolite description and the second column contains the corresponding numerical values and units. Finally, *IMT and FMD data* from each patient are entered to the system either manually through a GUI or through a tab delimited text file. If the second option is used, which data to use from within the file will also be supplied. In addition some calculated variables may be specified which will automatically be calculated from the data.

Data Analysis and **Visualization** modules were deployed on several client machines, distributed worldwide, and used to perform biostatistical analyses on the integrated and normalized data retrieved from the MK-DCNS. A typical analysis task is the following one:

1. the user imports the data;
2. the user partitions the data according to one of the phenotypes, *i.e.* all people with BMI below XXX are in class A and all people with a BMI above XXX are in class B;
3. with the above partition the user runs a differential expression analysis to see which genes are differentially expressed in the 2 groups;
4. the user runs a *GO (Gene Ontology)* analysis to see which GO terms are enriched in the above ranked list of genes;
5. results are graphically visualized.

Other types of analysis which can be performed are *classification*, *clustering*, *class discovery*, and *sequence motifs finding*.

The second instance of the pilot experiment will require further recruitment, up to a estimated sample of about 150-200 subjects. This new study will be performed to test an enhanced MK IT system prototype, including the following modules:

- the **Reporting** module, providing two web interfaces respectively for the definition of report structures, and for the creation and publication of useful reports after each data analysis experiment step;

– the **Workflow Management** module, allowing to define and control of the experiments.

All modules have been or are being developed by different IT partners of the Multi-Knowledge project. System integration is facilitated by the adoption of Web Service technologies. One important aspect is security, including user authentication and authorization (with a role-based policy) and sample data anonymization and protection.

5 Preliminary Results

The implemented components of the MK system have been used to perform preliminary statistical analyses on the first MK pilot experiment. All the major statistical scenarios have been tested.

As first step, we tried to reproduce results already known from the medical literature, such as the difference in gene expression profile between the two genders. The major differences between males and females in already published studies are related to the expression of genes codified in the sexual chromosomes, being Y-chromosome genes expressed exclusively in men and X-linked genes significantly more expressed in women. Using the MK overabundance analysis algorithm [7], we identified the list of genes differentially expressed between the two genders. As shown in figure 3, the algorithm discriminated almost perfectly (only one misplacement) between the two genders and the differentially expressed genes driving this difference were several gender-related genes (especially genes from the Y chromosome), as it is evident from the heatmap representation of the 50 most differentially expressed.

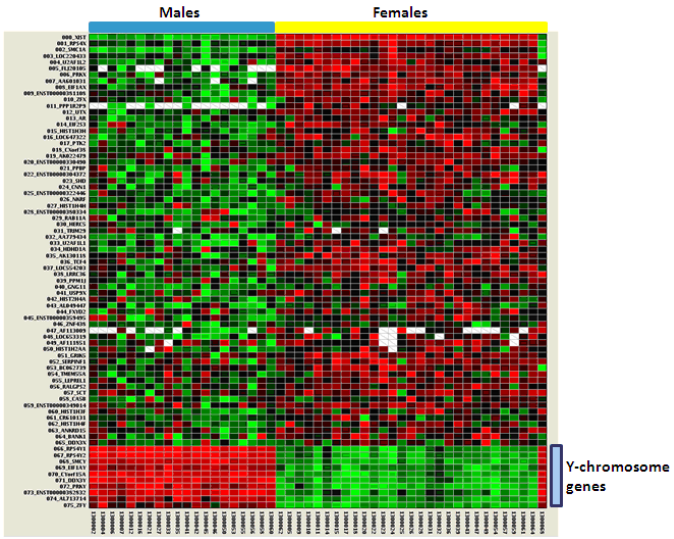

Fig. 3. MK overabundance analysis algorithm results.

As next step, we analyzed the differential expression between two groups of patients selected on the basis of a clinical partition. We chose smoking habit as a clinically relevant variable with a natural dichotomous distribution. We identified two groups of individuals (current smokers and never smokers) and we explored the differential expression between the two groups using partitioning function and overabundance algorithm implemented in the MK system.

The results of the analysis show a significant difference in expression profile of several genes between smokers and non-smokers, and the subsequent GO analysis identifies GO terms related to inflammatory response (such as ”defense response” and ”immune response”) as the most over-expressed in smokers. Results were also confirmed using external, validated GO analysis tools (like EASE [8]). On the overall population in study we also analyzed quantitative parameters of smoking, such as number of smoked cigarettes per-week (cig/wk). To identify the profile of differentially expressed genes between null to light smokers and heavy smokers, we run a partition search through a novel MK algorithm performing TNoM overabundance analysis [9] in all the possible partitions of the selected variable. The algorithm identified the two cutoff values for cigarettes/week to have the best discrimination between the two groups as less than 35 and more than 126 cig/wk, meaning about less than 5 cigarettes and almost 1 pack per day. Figure 4 illustrates the heatmap representation of the genes most differentially expressed in the two groups showing an extremely good partition between the two groups.

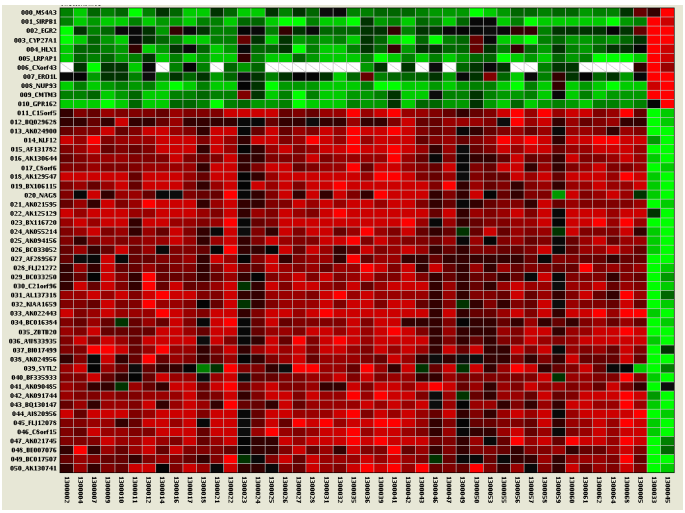

Fig. 4. Difference in expression profile of several genes between smokers and non-smokers.

Significant results in terms of differential expression were also found for several other clinically relevant variables through bi-partition discovery analysis, including plasma LDL cholesterol concentration (the major risk factor for cardiovascular dis-

ease), hs-CRP (the major biomarker for atherosclerosis-related inflammation), and IMT (a surrogate markers for early vascular signs of atherosclerosis). In case of IMT and hs-CRP, the cut-off thresholds identified by the partition algorithm as the best threshold to classify the sample in two groups were 1 mm and 3 mg/L that are well known values used in clinical research to identify subjects with high degree of inflammation and vascular wall damage respectively.

The GO analysis operated on the differentially expressed genes in subjects with high LDL cholesterol compared to low LDL cholesterol showed a significant enrichment of several GO terms related to inflammation and metabolism, as shown in figure 5 using the MK visualization module.

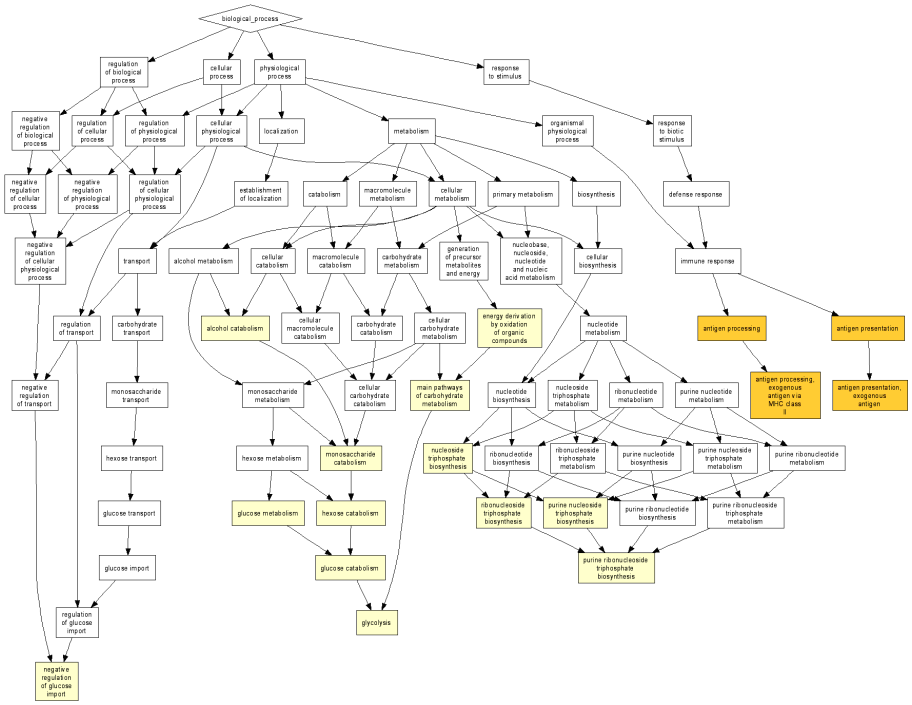

Fig. 5. GO analysis results.

6 Conclusions and Future Work

In this paper we described the collaborative environment proposed by the Multi-Knowledge project. We illustrated the generic template of Multi-Knowledge workflow processes, and the pilot experiment which we are carrying out. We presented the first Multi-Knowledge IT platform prototype, along with preliminary scientific results obtained from the analysis of microarray/clinical data of 50 subjects.

During the second year of the project, the Workflow Engine module and the Reporting module will be developed and integrated in the prototype. The pilot experiment will be completed considering 100 more samples.

References

1. Multi-Knowledge Consortium: Multi-Knowledge home page. http://www.multiknowledge.eu
2. Cocoon consortium: COCOON EU Project homepage. http://www.cocoon-health.com
3. Artemis Consortium: ARTEMIS EU Project homepage. http://www.srdc.metu.edu.tr/webpage/projects/artemis/index.html
4. Hull, D.: The Biological Web Services page. http://taverna.sourceforge.net/index.php?doc=services.html
5. Liefeld, T. and Reich, M. and Gould, J. and Zhang, P. and Tamayo, P. and Mesirov, J. P.: GeneCruiser: a Web Service for the annotation of microarray data. Bioinformatics **18** (2005) 3681–3682
6. Krishnan, S. and Baldridge, K. and Greenberg, J. and Stearn, B. and Bhatia, K. An End-to-end Web Services-based Infrastructure for Biomedical Applications. 6th IEEE/ACM International Workshop on Grid Computing, Seattle, Washington, USA, November 2005.
7. Ben-Dor, A. and Friedman, N. and Yakhini Z. Overabundance Analysis and Class Discovery in Gene Expression Data. Technical Report 2002-50, School of Computer Science & Engineering, Hebrew University , 2002.
8. EASE: the Expression Analysis Systematic Explorer. http://david.abcc.ncifcrf.gov/ease/ease.jsp
9. Ben-Dor, A. and Bruhn, L. and Friedman, N. and Nachman, I. and Schummer, M. and Yakhini, Z. Tissue classification with gene expression profiles. Journal of Comput. Biol. 2000; 7(3-4):559-83.