# Advancing Clinico-Genomic Research Trials via Integrated Knowledge Discovery Operations

George POTAMIAS[a,1], Manolis TSIKNAKIS[a], Vaggelis PAPOUTSIDIS[a], Alexandros KANTERAKIS[a], Kostas MARIAS[a], and Dimitris KAFETZOPOULOS[b]

[a] *Institute of Computer Science, FORTH, Heraklion, Crete, Greece*
[b] *Institute of Molecular Biology and Biotechnology, FORTH, Heraklion, Crete, Greece*

**Abstract.** In this paper we present the results on R&D efforts focusing on the development of an integrated GRID-enabled platform - founded on the Globus Toolkit, to enable efficient and reliable combined clinico-genomic knowledge discovery (CGKD) processes. The whole endeavor is considered in the context of biomedical informatics research and aims towards the realization of an Integrated and GRID-enabled Biomedical Infrastructure (IGBI). The presented CGKD scenario, and its process realization is inspired and implements a multi-strategy data-mining approach that seamlessly integrates three distinct data-mining components, namely: clustering, association rules mining, and feature-selection. The experimental platform is applied on a real-world microarray domain and respective study (gene-expression profiling and clinical outcome of breast cancer patients). Assessment of the overall methodology and of the results demonstrates the rationality and reliability of the approach.

## 1. Introduction

Recent advances in post-genomics research have resulted in an explosion of information, data and knowledge about major diseases, such as cancer, and their treatment. As a result, the application of related technologies to the study of diseases is slowly shifting to the analysis of clinically relevant samples such as fresh biopsy specimens and fluids. The respective scientific and technological challenges push for *trans-disciplinary* team science and *translational* research as the means to bring basic discoveries closer to the bedside [10]. In this context the design, development and delivery of up-to-date methods, systems and tools to support *discovery driven translational research* is inevitable. This task is comprised in the research agenda put forward by the scientific discipline of *Biomedical Informatics* – BMI [2], [9], also realised by various EU projects [1], [4], [15]. BMI melds the study of biomedical computer science with analyses of biomedical information and knowledge, thereby addressing specifically the interface between computer science and biomedical science [16].

Up to now, the lack of a common infrastructure has prevented researchers from being able to mine and analyze disparate data sources, due to the absence of a uniform platform supporting the seamless integration and analysis of both clinical and genomic data. To cope with this problem, we present the basics of an *Integrated GRID-enabled Biomedical Infrastructure* (IGBI) as approached by *ACGT* - a new EU/IST funded integrated project. Data mining & knowledge discovery operations are central to this infrastructure. In this context, and towards meeting the challenge of discovery driven translational research, we define and present an integrated *Clinico-Genomic Knowledge Discovery* (CGKD) process scenario, suited for the linkage and processing of both patients' gene-expression (microarray) and clinical data. The whole approach composes a 'screening' scenario for the careful identification of those patient cases and genes, which are more suitable to feed a gene-selection process. The proposed process is inspired by respective multi-strategy data mining and case-based reasoning methodologies [6]: *Clustering* – based on a categorical *k-means* clustering algorithm to induce gene clusters that best describes the available patient cases are selected; *Association Rules Mining* – to discover 'causal' relations between genes and patients' clinical features; and *Feature Selection* – to select those genes that best discriminate between patients' classes (i.e., disease state, survival category etc). The work reported in this paper was carried-through in the context, and with the support of two projects namely, *PrognoChip* [5], [11], and INFOBIOMED [4]. In the next section we present the basics of IGBI. In section 3 the integrated CGKD approach is presented in detail. Section four covers preliminary experimental results on a real-world case study. Finally, in section five we conclude and point to future R&D directions.

[1] Corresponding Author: George Potamias, FORTH-ICS, Vassilika Vouton, 71110 Heraklion, Crete, Greece. E-mail: potamias@ics.forth.gr

## 2. Towards an Integrated GRID-enabled Biomedical Infrastructure

Grid computing enables the virtualization of distributed computing over a network of heterogeneous resources giving users and applications seamless, on demand access to vast IT capabilities [3]. Grid computing provides a novel approach to harnessing distributed resources, including applications, computing platforms or databases and file systems.

Within such a context, the implementation of the EU funded Integrated Project named "*Advancing Clinico-Genomic Trials on Cancer: Open Grid Services for Improving Medical Knowledge Discovery*" – *ACGT* is just launched. The ultimate objective of the *ACGT* project is the provision of a unified technological infrastructure which will facilitate the seamless and secure access and analysis of *multi-level* clinico-genomic data, enriched with high-performing knowledge discovery operations and services (see Fig. 1). In so doing, *ACGT* aims to contribute to the realisation of the *genomic* and *individualised medicine* vision [14] towards: (a) the advancement of cancer research for revealing the influence of genetic variation in oncogenesis, (b) the promotion of molecular classification of cancer and the development of individualised therapies, and (c) the development of realistic and reliable *in-silico* tumour growth and therapy response models (for the avoidance of expensive and often dangerous examinations and trials on patients).

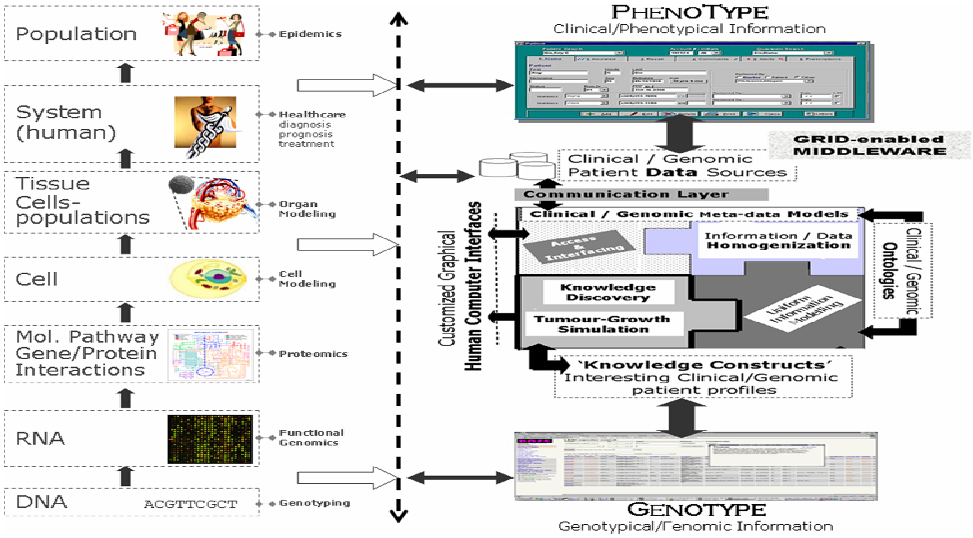



**Fig. 1:** The envisioned ACGT GRID-enabled infrastructure and integrated environment – integration to be achieved at all levels, from the molecular to system and to the population

The work in ACGT targets to the delivery of an *Integrated GRID-enabled Biomedical Infrastructure* (IGBI) that comprises the following components:

- *Biomedical technology Grid layer* - the basic "*Grid engine*" for the scheduling and brokering of resources;
- *Distributed Data Access and Applications* - to provide seamless and interoperable (ontology-based) access-to and retrieval-of distributed data sources;
- *Data Mining and Knowledge Discovery* tools – intelligent data analysis services based on the deployment of *open*, *interoperable* and GRID-enabled data mining and analysis software tools and services [8] to support knowledge discovery from combined clinico-genomic biomedical data;
- *Ontologies and Semantic Mediation* tools - building on the various ontologies and controlled vocabularies that have grown over the years for providing a shared language for the communication of biomedical information, e.g., Gene Ontology (www.geneontology.org), MGED Ontology (www.mged.org), UMLS (www.nlm.nih.gov/research/umls/) etc;
- *In-silico oncology tools – to* demonstrate the added value of in-silico modelling of tumour growth (for the case of cancer disease) and therapy response; and
- *The integrated ACGT environment* – assembly of tools and services based on complex *workflows*.

## 3. Integrated Clinico-Genomic Knowledge Discovery: GRID platform and Process Scenario Implementation

### *3.1. The GRID-enabled platform*

The work reported in this paper was (mainly) carried-through in the context of *PrognoChip* – a multi-disciplinary research project [5], [11]. In developing the *PrognoChip* clinico-genomic environment, it was decided to explore a service oriented architecture and to study the potential benefits of using Grid technology, building upon the results of key health-related GRID projects, such as caBIG (https://cabig.nci.nih.gov), BIRN (http://www.nbirn.net), MEDIGRID (http://www.creatis.insa-lyon.fr/MEDIGRID/), MyGRID (http://www.mygrid.org.uk) and others. DiscoverySpace (http://www.bcgsc.ca/discoveryspace/). We selected the *Globus Toolkit* for the implementation of the Grid middleware and building our open Grid layer. The Globus Toolkit is an open source software toolkit developed by the Globus Alliance and many others (www.globus.org). It provides Grid services that meet the requirements of the Open Grid Service Architecture and are implemented on top of the Web Service Resource Framework. It includes software for security, information infrastructure, resource management, data management, communication, fault detection, and portability. The most important components of the Globus Toolkit involved in our implementation are (see Fig. 2):

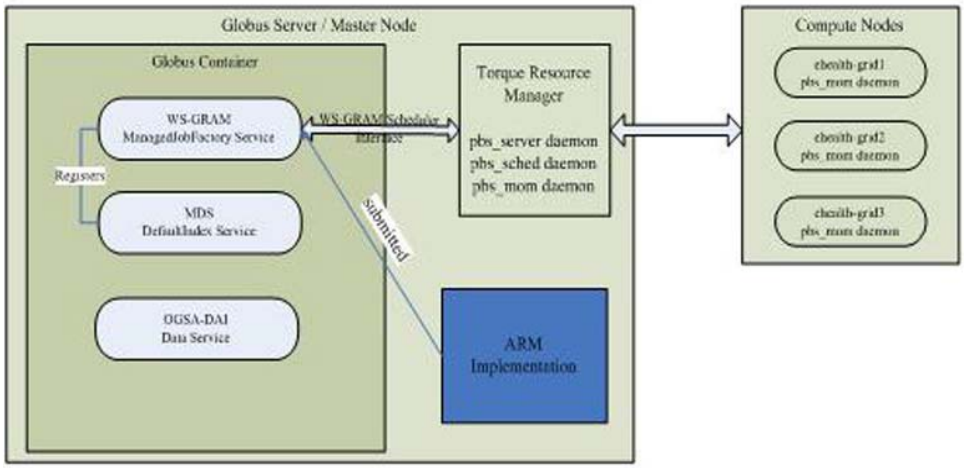



**Fig. 2:** The components of the Grid system

*Resource Allocation and Management (WS-GRAM)***.** The component of Globus Toolkit 4 responsible for allocating jobs to resources that we have already set up and governed by local schedulers. Integration of WS-GRAM component with local schedulers allows utilisation of resources in a uniform way, and is based on open protocols.

*Service Monitoring and Discovery.* In a Grid environment, the set of resources available for use by a virtual organization can change frequently. The Globus Toolkit's solution to these closely related problems is its Monitoring and Discovery System (MDS) [18]. MDS4 is a suite of web services to monitor and discover resources and services on Grids. This system allows users to discover what resources are considered part of a Virtual Organization (VO) and to monitor those resources. MDS4 includes two services: an *Index Service*, which collects data from various sources and provides a query/subscription interface to that data, and a *Trigger Service* which collects data from various sources and can be configured to take action based on that data.

*Access to and heterogeneous data integration.* OGSA-DAI (www.ogsadai.org.uk) is a middleware product that allows data resources to be accessed via web services. An OGSA-DAI web service allows data to be queried, updated, transformed and delivered. These services can be deployed within a Grid environment, thereby OGSA-DAI provides the means for users to Grid-enable their data resources. OGSA-DAI expose data resources through services. Clients interact indirectly with data resources via these services. Each data service exposes zero or more data service resources. A data service resource is the component of OGSA-DAI that through a data resource accessor provides direct access to a single data source (database).

*The Grid Security Infrastructure (GSI).* The portion of the Globus Toolkit that provides the fundamental security services needed to support Grids. Security is one of the most important parts of a Biomedical Grid, since a grid implies crossing organizational boundaries. GSI is composed of a set of command-line tools to manage certificates, and a set of Java classes to easily integrate security into web services.

*3.2. Integrated Clinico-Genomics Knowledge Discovery: Process Scenario Implementation*

The integrated *clinico-genomic knowledge discovery* (CGKD) process scenario is suited for linkage of patients' gene-expression (microarray) and clinical data. The whole approach composes a '*screening*' scenario for the careful identification of those patient cases and genes, which are more suitable to feed a gene-selection process. The proposed process is inspired by respective *multi-strategy* data mining and machine learning methodologies. The approach is based on the smooth integration of three distinct data-mining methodologies: (**i**) *Clustering* - based on a novel *k-means* clustering algorithm operating on *categorical* data, named *discr-kmeans* [6], [13]. 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 'up'- or, 'down'-regulated gene-expression values (the '*strong*' clusters). The method is implemented as an integrated system for mining gene-expression (microarray) data – the MineGene system [7], [12]; (**ii**) *Association Rules Mining (ARM)* - for the discovery of strong 'causal' relations (rules with high confidence) between genes and patients' features, and operate on the genes and patient cases being covered by strong clusters; and (**iii**) *Feature Selection* - for the selection of the most discriminant genes, i.e., genes being able to distinguish between patients' pre-specified classes (disease state, survival category etc).

The details of the CGKD scenario are shown in Fig. 3. The figure shows the quest-specifics for each of the smoothly integrated data-mining methods as well as the respective findings with their utilization and interpretation.

- *Strong-clusters and their interpretation.* We are interested in 'strong' clusters because we want to identify potential subsets of samples that tend to exhibit mainly 'up'- or 'down'-regulated expression levels for the respective cluster's genes. This is why we decide to discretize the continuous gene-expression levels with a discretization value of n=2 (the implemented discr-kmeans allows more that two discretization values). This means that for these samples (the '*strong-samples*'), the respective cluster's genes tend to be 'dominantly' up- or down-regulated under the specifics of the experimental conditions. The genes of a cluster, accompanied by the 'strong-samples' covered by this cluster may be interpreted as a combined '*clinico-genomic feature*' linking patient cases and their genomic (gene-expression) profiles. The quest now is whether *causal* relationships between genomic and clinical profiles do exist.
- *Interpretation and utility of discovered association rules.* Each association rule contains a combination of clinico-genomic features that uncovers not only significant but also *causal* relations between genomic and clinical patient profiles. Each association rule is taken as a *medium to focus on the genes and patient cases covered by it.* The expert (molecular biologist or, physician) may inspect the discovered association rules and focus on the most interesting (i.e., highly confident) ones. Then, a gene-selection process is initiated and operates just on the sets of genes and patient cases being covered by the focused association rules, in order to identify genes that distinguish between patient case classes (e.g., "survival greater that 5 years" vs. "survival less than 5 years"; look at the next section).

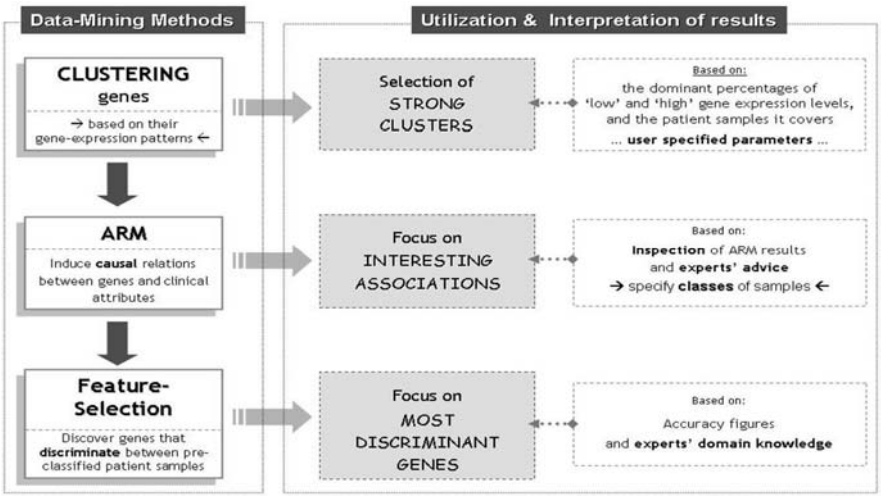



**Fig. 3:** Integrated Clinico-Genomic Knowledge Discovery scenario enabled by the smooth integration of different data-mining methods

## 4. A Real-World Case Study

We applied the presented integrated CGKD process scenario on a real world clinico-genomic domain. For the reference (publicly available) study with which we compare our findings see [17]. The data includes the gene-expression profiles of 24.481 genes over 78 breast-cancer patient samples; 44 of them with a status of over five years survival, and 34 with a status of less than five years survival. The clinical profiles of the patients are also provided. The clinical data refer to a number of features including: *age* of the patient; *Lymphocytic-Infiltration* status of the tumour; the *estrogen* and *progesterone* receptor's profiles of the patients, as well as their *prognostic status* (i.e., 'bad' or, 'good', for less and over five years survival, respectively).

In the experiment reported in this paper, we focus on the *prognostic* status, trying to discover reliable associations between the prognostic profiles of patients and their gene-expression background. The findings were as follows:

(a) *Clustering genes and selection of strong clusters.* For discr-kmeans clustering the number of clusters were pre-specified (90 clusters were requested) in order to adequately cover all input genes. In a next step we selected the clusters that show a strong relation with the respective clinical features (as described in the previous section). This process is (automatically) performed with respect to the following thresholds: #samples: ≥10, #genes: ≥100, with the respective threshold for considering a sample as strong for a cluster to have at least 90% 'up'-regulated genes in the respective cluster. We ended up with a set of 13 gene-clusters.

(b) *ARM and causal clinico-genomic relations.* In order to find informative and highly confident association rules we selected all the genomic features to participate in the 'IF' scope of the rule and the remaining clinical features to participate in the 'THEN' part of it. We set minsup=10, and minconf=70 (for minimum support and confidence of each rule, respectively), and focused on "*followup-time*" in years (as the clinical feature) on all 13 gene-clusters. In the resulting association rules only 3 out of the 13 gene clusters appeared, covering: 37 out of the input 78 samples, and 5936 genes (5503, 284 and 149 for the three respective clusters).

(c) *Gene-Selection.* Applying the feature-selection process (presented in the previous section) on the set of 37 cases and the set of 5936 genes we end-up with a set of 100 most-disciminant genes (column 2 in table 1, below, where the respective accuracy results are presented). The feature-selection process was also performed on all 78 patient-samples and on an independent test-set of 19 patient samples (columns 4 and 5 of table 1, respectively). The presented results are indicative for the rationality and reliability of the CGKD approach.

**Table 1.** Comparative gene-selection (accuracy) results after running the implemented CGKD scenario

| | **#SG*** | on **37** samples** | on **78** total samples | on **19** test samples*** |
|---|---|---|---|---|
| CGKD | 100 | 100% | 85.9% | 89.5% |
| Reference Study | 70 | NA[+] | 80.8% | 89.5% |

*SG: number of selected genes

**The samples selected by the CGKD process

***Independent set of samples (left-out during training)

[+]NA: Not Applicable

## 5. Conclusions

An GRID-enabled biomedical infrastructure is presented, as envisioned by ACGT - a newly funded EU/IST integrated project. In this context a preliminary implementation of a GRID-enabled integrated clinico-genomic knowledge discovery process scenario is presented. The process is based on: (i) a GRID-platform founded on the Globus Toolkit, and (ii) the smooth integration of different data-mining methodologies (clustering, association rules mining and feature-selection) as a means to link and '*screen*' between those patient cases and genes which are more suitable to feed a gene-selection process. The platform and knowledge-discovery process was implemented in the context of PrognoChip, multidisciplinary, research project.

The implemented process was applied on a real-world microarray domain concluding into strong 'causal' relations between specific groups of genes and patient cases' samples. Comparison results with a reference real-world case study are also presented. The implementation has proved the benefits of employing GRID technologies

and combination (in a multi-strategy mode) of data-mining methodologies for addressing the complex computational tasks found in the multidisciplinary domain of clinico-genomics and biomedical informatics.

Our future R&D plans focus on the design and implementation of a Web-based environment to support the presented CGKD process.

**Acknowledgments.** The reported work was partly supported by the following projects: PrognoChip (EPAN program, funded by the Greek Secretariat for Research & Technology), and INFOBIOMED (FP6/2002/IST-507585). Gratitude should be also given to the ACGT project (FP6/2004/IST-026996) consortium.

## References

[1] BIOPATTERN. Computational Intelligence for Biopattern analysis in Support of eHealthcare. Network of Excellence, EU/FP6/IST/NoE project [http://www.biopattern.org].

[2] Eich HP, de la Calle G, Diaz C, Boyer S, Pena AS, Loos BG, Ghazal P, Bernstein I. Practical Approaches to the Development of Biomedical Informatics: the INFOBIOMED Network of Excellence. Stud Health Technol Inform. 2005; 116:39-44.

[3] Foster I. The Grid: A New Infrastructure for 21st Century Science. Physics Today 2002, 55(2):42-47.

[4] INFOBIOMED. Biomedical informatics to support individualised healthcare. Network of Excellence, EU/FP6/IST/NoE project [http://www.infobiomed.net].

[5] Kafetzopoulos D. The Prognochip Project: Transcripromics and Biomedical Informatics for the Classification and Prognosis of Breast Cancer. ERCIM News 2005, 60, 27-28.

[6] Kanterakis A, Potamias G. Supporting Clinico-Genomic Knowledge Discovery: A Multi-Strategy Data Mining Process. 4th Hellenic Conference on Artificial Intelligence (SETN'06), LNAI (to appear).

[7] Kanterakis A. Gene Selection & Clustering Microarray data: The MineGene System. MSc thesis, Dept. of Computer Science, University of Crete, 2005. [http://www.csd.uoc.gr/~kantale/Kanterakis_THESIS.pdf].

[8] Kickinger G, Brezany P, Min Tjoa A, Hofer J. Grid Knowledge Discovery Processes and an Architecture for Their Composition. In: IASTED Conference 2004, Innsbruck, Austria, February 17-19.

[9] Martin-Sanchez F, Iakovidis I, et al. Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. J Biomed Inform. 2004, 37(1):30-42.

[10] Parks MR, Disis ML. Conflicts of interest in translational research. Journal of Translational Medicine 2004, 2:28:1-4.

[11] Potamias G, Analyti A., Kafetzopoulos D, Kafousi M, Margaritis T, Plexousakis D, Poirazi P, Reczko M, Tollis IG, Sanidas ME, Stathopoulos E, Tsiknakis M, Vassilaros S. Breast Cancer and Biomedical Informatics: The PrognoChip Project. IMACS 2005: Computer Science and Artificial Intelligence – Bioinfoamtics session (organizer: Casimir Kulikowski), 2005: 11-15, Paris, France.

[12] Potamias G, Koumakis L, Moustakis V. Gene Selection via Discretized Gene-Expression Profiles and Greedy Feature-Elimination. LNAI 2004, 3025: 256-266.

[13] San OM, Huynh V-N, Nakamori Y. An alternative extension of the k-means algorithm for clustering categorical data. Int. J. Appl. Math. Comput. Sci. 2004, 14(2):241–247.

[14] Sander C. Genomic Medicine and the Future of Health Care. (2000), Science 2002, 287(5460):1977-1978.

[15] Semantic Mining. Semantic interoperability and data mining in biomedicine. Network of Excellence, EU/FP6/IST/NoE project [http://www.semanticmining.org/].

[16] Shortliffe EH. Biomedical Informatics: The Nature of the Discipline. Department of Biomedical Informatics, *Columbia University*, 2004 [http://www.dbmi.columbia.edu/about/definition/ definition.html].

[17] van't Veer L, et al., Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002, 415:530–536.

[18] Zhang X, Schopf J. Performance Analysis of the Globus Toolkit Monitoring and Discovery Service, MDS2. (2004), Proc. of the International Workshop on Middleware Performance (MP 2004), part of the 23rd International Performance Computing and Communications Conference (IPCCC).