Grid-based Knowledge Discovery in Clinico-Genomic Data

Michael May[1], George Potamias[2], and Stefan Rüping[1]

[1] Fraunhofer AIS, Schloss Birlinghoven, 53754 St. Augustin, Germany,
{michael.may,stefan.rueping}@ais.fraunhofer.de,
WWW home page: http://www.ais.fraunhofer.de
[2] Institute of Computer Science, FORTH, Heraklion, Crete, Greece
potamias@ics.forth.gr,
WWW home page: http://www.ics.forth.gr

Abstract. Knowledge discovery in clinico-genomic data is a task that requires to integrate not only highly heterogeneous kinds of data, but also the requirements and interests of very different user groups. Technologies of grid computing promise to be an effective tool to combine all these requirements into a single architecture. In this paper, we describe scenarios and future research directions related to grid-based knowledge discovery in clinico-genomic data, and introduce the approach taken by the recently launched ACGT project[3]. 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. The presented integrated clinico-genomics knowledge discovery (ICGKD) scenario and its process realization is based on a multi-strategy data-mining approach that seamlessly integrates three distinct data-mining components: clustering, association rules mining, and feature-selection. Preliminary experimental results are indicative of the rational 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 [1]. In this context the design, development and delivery of up-to-date methods, systems and tools to support knowledge discovery in clinico-genomic data is of major importance. This task is comprised in the research agenda put forward by the scientific discipline of Biomedical Informatics BMI [2], also realized by various EU projects,

[3] http://www.eu-acgt-org

e.g. INFOBIOMED[4], and the recently launcehd ACGT[5] project. 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.

The effective and efficient management and use of stored data, and in particular the transformation of these data into information and knowledge, is a key requirement for success in the biomedical domain. Knowledge Discovery (KD, aka Data Mining) is the de-facto technology addressing this information need. Data mining technology is used for the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. However, this field has mainly been concerned with small to moderately sized data sets, knowledge-weak domains, and within the contest of largely homogeneous and localized computing environments. These assumptions are increasingly not met in modern scientific environments. The shift to large-scale distributed computing has profound implications in terms of the way data are analyzed. Future data mining applications will need to operate on massive data sets and against the backdrop of complex domain knowledge. The domain knowledge (computer-based and human-based), the data sets themselves, and the programs for processing, analyzing, evaluating, and visualizing the data, and other relevant resources will increasingly reside at geographically distributed sites on heterogeneous infrastructures and platforms. Grid computing [3] promises to become an essential technology capable of addressing the changing computing requirements of future distributed knowledge discovery environments. Currently, several project such as DataMiningGrid[6] [4] and SIMDAT[7] are concerned with merging generic knowledge discovery services with grid technologies.

In this paper, we will review recent developments on grid-based biomedical research and point out future directions to facilitate an integrated, knowledge-rich, and highly performant analysis of clinical and genomic data. With respect to the last point, we present an Integrated Clinico-Genomics Knowledge Discovery (ICGKD) scenario and its realization be a multi-strategy data-mining process that smoothly integrates three data-mining approaches: clustering, association rules mining, and feature-selection.

The rest of the paper is organized as follows: the next section will introduce the main challenges of KD in clinico-genomic data, while Section 3 introduces grid-enabled Knowledge Discovery. In Section 4 we will present novel research directions targeted at solving the biomedical challenges using grid-based KD, in particular the approach taken by the ACGT project. Section 5 gives an example by describing the ICGKD scenario and its realization. In Section 6 a real-world case study is presented. Finally we conclude in Section 7.

[4] http://www.infobiomed.org
[5] http://eu-acgt.org
[6] http://www.datamininggrid.org/
[7] http://www.scai.fraunhofer.de/simdat.html

2 Challenges of Knowledge Discovery in Clinico-Genomics

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 may contain only about 100 examples. At the same time, the same data sets consists 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 knowledge discovery 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 (diagnosis, treatments, vital signs) 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. 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 [5] 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.

3 Grid-enabled Knowledge Discovery

Grid computing [3] 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. However, since the inception of Grid technology, the term resource has evolved to cover a wide spectrum of concepts. Standard grid solutions provide services such as job execution and monitoring, parallelization, distributed data access and security.

3.1 Data Mining on the Grid

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 standardization. In particular, in the application of models to unseen data, this data may come from very different sites than the original data set that was used for learning.
- 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, data mining algorithms consist of a set of almost identical, independent tasks, e.g. for parameter sweeps, cross-validation, 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.

All these tasks by themselves are not very complex and 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.

An approach to provide standard data mining services across a grid infrastructure is the integration of standard statistical and data mining toolkits, such as Weka [6] or R [7]. This approach was followed by the DataMiningGrid project, which allows to integrate Weka operators into a overall grid workflow. Figure 1 presents such an workflow. These workflows are described in the form of an XML document, which is executed by the main grid engine.

4 The ACGT approach to grid-based clinico-genomics knowledge discovery

The recently started ACGT project aims at:

- the delivery of a European biomedical grid infrastructure offering seamless mediation services for sharing data and data-processing methods and tools, and advanced security;

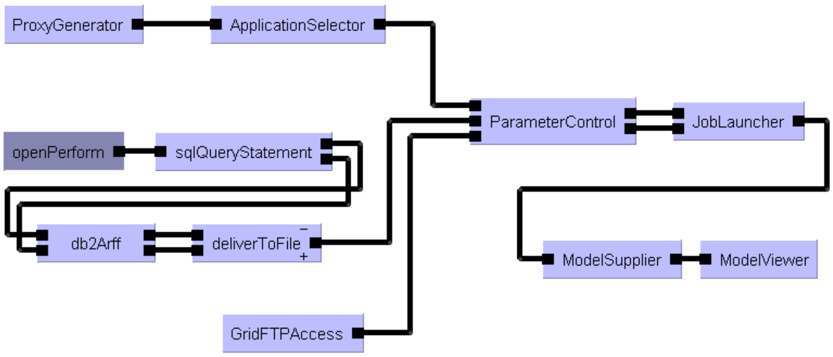


Fig. 1. An integrated grid and data mining workflow, including the selection of a data mining operator (upper left hand side), selection of distributed data (lower left), job execution (upper right) and display of the results (lower right).

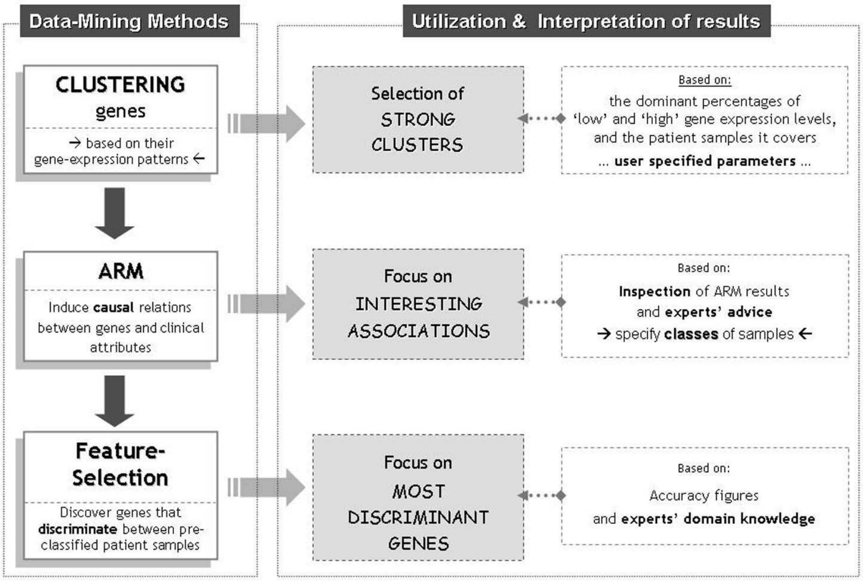
- the semantic, ontology based integration of clinical and genomic/proteomic information and data;
- the delivery of data-mining grid services in order to support and improve complex knowledge discovery processes.

The main challenge from the knowledge discovery side of the 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 other researcher. The efficient management of the different views and expertise of clinicians, biologists and data miners will be crucial. For the underlying principles of ACGT and the way it copes with these issues see [8]. In particular, we propose the use of the following techniques

- An ontology-based description of the application domain- taking into account standard clinical and genomic ontologies, nomenclatures and metadata, in order to retain semantic on all steps of the analysis and to guide the construction of data mining workflows.
- A matching ontology to describe data mining tasks and operators, including support to correctly translate research questions of the application domain into specific data mining tasks.
- A database of workflows plus appropriate meta-information to serve as a case base for selecting promising candidate workflows from similar tasks.
- The use of text-mining techniques to extract relevant knowledge from published papers. Text mining can also be used to connect research questions to data mining tasks and algorithms in a more freely structured way, when workflow descriptions and corresponding paper abstracts are joined into one document and are added to the available document corpus.
- The massive parallel use of data mining algorithms to search for dependencies in the cases where no prior expert knowledge is available.

5 An Exemplary Scenario

In order to exemplify our approach, let us discuss an Integrated Clinico-Genomic Knowledge Discovery (ICGKD) scenario that has been discussed in [9]. The scenario is depicted in Figure 2 and consists of three steps, first the clustering of genes based on their gene-expression patterns in order to identify potentially useful subsets of genes, the discovery of association rules to discover causal relations between genes and clinical attributes (in this case the prognostic status for breast cancer patients), and post-processing by feature selection in order to focus on the most discriminating genes, based on both accuracy and experts' domain knowledge.



Data-Mining Methods
Utilization & Interpretation of results
CLUSTERING genes
→ based on their gene-expression patterns ←
Selection of STRONG CLUSTERS
Based on:
the dominant percentages of 'low' and 'high' gene expression levels, and the patient samples it covers
... user specified parameters ...
ARM
Induce causal relations between genes and clinical attributes
Focus on INTERESTING ASSOCIATIONS
Based on:
Inspection of ARM results and experts' advice
→ specify classes of samples ←
Feature-Selection
Discover genes that discriminate between pre-classified patient samples
Focus on MOST DISCRIMINANT GENES
Based on:
Accuracy figures and experts' domain knowledge

Fig. 2. Integrated Clinico-Genomic Knowledge Discovery Scenario

Using the approaches described in Section 4, this scenario can be extended and realized in multiple ways

- A researcher looking for novel ways to analyze some type of cancer that is known to be related to breast cancer could find the workflow described here by using an ontology-based query to the workflow case base. Executing this workflow on the new set of data would only need to replace the data selection part of the workflow.
- The workflow designer could be interested in improving the analysis by comparing different kinds of *clustering* algorithms in order to identify clusters of genes that exhibit significant relations with specific patient samples. Most

standard data mining toolkits already come with a large variety of clustering algorithms, and using a grid environment the parallel execution and comparison of all these alternatives is very easy. In the real-world application presented in the next section the clustering operation is realized by a novel *k-means* clustering algorithm that operates on discretised gene-expression data.

- Next the researcher may be interested to identify 'strong' associations (i.e., utilizing a frequent itemset discovery algorithm) between particular patients' clinical profiles and features with the gene-clusters discovered in the second step. The developer of the association rule algorithm could be interested in developing a parallel version to speed up the computation. This task is very much alleviated by the use of standard components in a grid-aware toolkit. For the real-world experiment presented in the next section a grid-enabled association rules mining system is utilized, the *HealthObs* system [10]
- Focusing on the genes and samples covered by discovered *high-confident* association rules the researcher may be then interested to identify potential *gene-markers*, i.e., genes that best *discriminate* between specific patient status (e.g., good vs. bad prognosis). Particular parallelized and grid-enabled feature-selection and classification algorithms from standard data-mining toolkits could be utilized. The real-world experimental application in the next section utilizes a particular gene-selection system, the MineGene system [11]

5.1 Towards Strong Clinico-Genomic Profiles

A lot of work has been done in identifying co-regulated groups or, clusters of genes [12], clusters of patient samples [13], discriminant set of genes [14], and methods to reduce the dimensionality and complexity of gene-expression data [15]. In this paper we introduce and utilize a discretized two-dimensional k-means clustering algorithm, named *discr-kmeans*, that primary identifies clusters of co-regulated genes. The algorithm resembles similar approaches presented in [16] and [17], and further exemplified in [18]. With a subsequent filtering approach the gene clusters that exhibit, in an adequate number of samples, 'strong' gene-expression profiles are selected. A sample with a strong gene-expression profile for a specific cluster of genes is one that exhibits *dominantly* 'high' or, 'low' gene expression levels. The adequate number of samples, as well as the percentage for considering a sample's expression profile as strong is set by the user.

Assume s samples, g genes, and a 2-dimensional matrix, $M(s \times g)$ that holds the respective gene-expression matrix. The discr-kmeans algorithmic process unfolds into two steps:

- Step-1: *Discretization.* We proceed with a method to overcome the error-prone variance of gene-expression levels by discretizing the respective continuous gene-expression values. A gene-expression value may be assigned to an (ordered) nominal value; assume n such ordered values. In the case of

$n = 2$, value 1 is interpreted as of *low*, and value 2 as of *high* expression level. Define,

$$w_i = \frac{max(g_i) - min(g_i)}{n}$$

where, $max(g_i)$ and $min(g_i)$ the maximum and minimum expression values of gene g_i, respectively. The discretized transform, $V_d(s_j, g_i)$, of gene's g_i continuous value, $V(s_j, g_i)$ in sample s_j is computed by:

$$V_d(s_j, g_i) = \begin{cases} n & \text{if } V(s_j, g_i) = max(g_i) \\ \lfloor \frac{V(s_j, g_i) - min(g_i)}{w_i} \rfloor + 1 & \text{otherwise} \end{cases}$$

where, $\lfloor fraction \rfloor$ the integer part of the fraction. It can be easily checked that the computed gene's discretized values range from 1 to n.

- Step-2: *Clustering.* The main difference between normal k-means and discr-kmeans is that each cluster's center is not represented by the average value of the cluster's genes values but, by a 2-dimensional matrix that contains the percentage of the discretized cluster's gene values, $C_k(s, n)$. For a sample s_j, and $p \in [1, n]$, $C_k(s_j, p)$ is the percentage of genes in cluster k, the discretized expression-values of which, with respect to sample s_j, is p. For example, in a domain with three samples and discretization value $n = 2$, a cluster's center is an array like the following:

Value →	1 (*low*)	2 (*high*)
Sample ↓	%	%
1	80	20
2	55	45
3	10	90

In the above example matrix, 80% of all the genes in the cluster exhibit low (discretized value = 1) expression levels in sample 1. Analogously, 90% of the cluster genes exhibit a high expression profile (discretized value = 2) in sample 3. With appropriate (user defined) thresholds, these profiles are assumed to *dominate* the respective samples. In other words, the induced cluster of genes seems to be linked and correlated with dominantly low, or high expression profiles for the specific samples. Clustering unfolds into the standard k-means iterative process. A basic difference is the way that the distance of a gene and the center of a cluster is computed. The distance between a cluster $C_k(s, n)$ and a gene g_i is computed as:

$$dist(C_k, g_i) = \sum_{l=1}^{s} C_k(s_l, V_d(s_l, g_i))$$

After clustering process converges, we end-up not only with gene clusters but with an indication of how 'strong' a cluster is. A cluster is considered as 'strong' if it exhibits dominant discretized value percentages in an adequate number of samples, i.e, close to 0%, or close to 100%. We are interested in 'strong' clusters because we want to identify potential subsets of samples that tend to exhibit mainly dominantly high or low expression levels for the respective genes in a cluster. This is why we decide to discretize the continuous gene-expression levels with a discretization value of $n = 2$. For these samples – referred as *strong samples*, the respective cluster's genes tend to be dominantly up- or, down-regulated. The genes of a cluster, accompanied by their respective strong samples may be interpreted as a combined *clinico-genomic attribute* linking patient cases and their genomic (gene-expression) profiles. The quest now is about the *causal* relations that hold between such genomic and clinical profiles.

5.2 Causal Relations in Clincio-Genomic Profiles

In the present study we utilized HealthObs, a system that incorporates Association Rules Mining (ARM) operations specially suited for the clinical domain [10]). HealthObs is able to operate over an integrated electronic health care record environment [19]. The special services that HealthObs brings relate to: (i) ease in query formulation, via a friendly GUI interface- flexible enough to enhance the naturalness of data exploration inquiries; (ii) imposition and utilization of ARM operations directly on-top of XML structures (instead of flat files or, specific databases); and (iii) friendly visualization operations that ease inspection, filtering and interpretation of the discovered association rules.

5.3 Selecting Discriminatory Genes via Feature-Selection

For the case of the clinico-genomic inquiry and exploration process, each association rule may be 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 ones that seem more interesting for the scope of the inquiry. Then, a *gene-selection* process may be called to operate just on the sets of genes and patient cases being covered by the focused association rules. Provided that a specific clinical feature of interest is targeted (e.g., 'survival over 5 years' vs. 'survival less than 5 years') particular gene-markers may be identified. In the present study we utilize a *feature-selection* method specially suited for the task of selecting discriminant genes, i.e., set of genes able to distinguish between particular pres-classified patient samples. A detailed description of the method may be found in [11]. The method is implemented in an integrated system for mining gene-expression (microarray) data - the MineGene[8] system. The method is composed by three components: discretization of gene-expression data, ranking of genes, and greedy feature-elimination (or, addition) accompanied with a classification metric to predict patient class categories (e.g., clinical outcome).

[8] http://dlib.libh.uoc.gr/Dienst/Repository/2.0/Body/uch.csd.msc/2005kanterakis/pdf

6 A Real-World Application

We applied the presented ICGKD scenario on a real world clinico-genomic domain. For the reference study (accompanied with public available data[9]) with which we compare our findings see [20]. 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 profiles* of the patients, as well as their prognostic status - *bad* or, *good*, for less and over five years survival, respectively. In the reported experiment we focus on the clinical-outcome feature, trying to discover reliable associations between the prognostic profiles of patients and their gene-expression background. Experimental set-up and findings follow.

- *Clustering genes and selection of strong clusters.* For discr-kmeans clustering operation the requested number of clusters was set to 90 (so that all input genes are appropriately covered). In order to consider a cluster of genes as strong (i.e., set of genes which exhibit dominantly up- or, down-regulated profiles in an adequate number of samples) the following parameters were set: $minimum - number - of - genes \geq 100$; $minimum - number - of - samples - with - a - dominant - genes - profile \geq 10$; and $percentage - of - genes - with - dominant - genes - profile - per - sample \geq 90\%$ of up- or, under-regulated genes in the respective cluster. We ended up with a set of 13 gene-clusters.
- *Association rules mining 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 *follow-up* (the clinical-outcome or, prognosis) feature to participate in the *THEN* part of them (a service offered by the HealthObs system). We set $minsup = 10$, and $minconf = 70$ for the minimum support and confidence of each rule, respectively. In the resulting association rules only 3 out of the 13 gene clusters appeared, covering 37 (from a total of 78 input) samples, and 5936 genes (5503, 284 and 149 for the three respective clusters).
- *Gene selection.* Applying the gene-selection process of MineGene on the set of 37 sample cases, and the set of 5936 genes we end-up with a set of 100 most-discriminant genes that exhibit an (fitness) accuracy figure of 100% (column 2 and 3 in Table 1, respectively). The gene-selection process was also performed on all 78 patient-samples, as well as on an independent test-set of 19 patient samples (columns 4 and 5 in Table 1, respectively). The presented results are indicative for the rational and reliability of the ICGKD approach.

[9] http://www.rii.com/publications/2002/vantveer.html

Table 1. Comparative accuracy results (for gene-selection) after running the presented ICGKD process (*SG: number of selected genes, **samples selected by the ICGKD process, ***NA: not applicable).

	#SG*	**37** samples**	**78** total samples	**19** test samples
ICGKD	100	100%	85.9%	89.5%
Reference study	70	NA***	80.8%	89.5%

7 Conclusions

Recent advances in post-genomics and especially in high-throughput technology (e.g., microarrays) offer the means to examine and profile the expression of all human genes and relate them with patients' disease profiles. It is an effective approach for developing disease prognostics expected to result into the identification of strong candidate targets for diagnosis and therapeutic intervention. The inherited huge amount of genomic information and respective patients' data calls for advanced data-mining tools and respective high-performing environments.

Grid technology promises to be an effective way to easily combine existing, previously independent approaches for knowledge discovery in clinico-genomic data into a single framework. The recently launched integrated project ACGT aims towards this direction. The provisioned technological platform will be validated in a concrete setting of advanced clinical trials on Cancer.

In this setting solutions to the problem of reducing the dimensionality of the search space, i.e., from thousands of genes to the most disease-status discriminant ones, are crucial in order to cope with the intrinsic noise and deliver reliable diagnostic and prognostic molecular/gene-markers. This is the target of the presented clinico-genomic knowledge discovery scenario and its realization via the smooth integration of different data-mining, namely: *which patients' clinical profiles relate and how with their respective genomic background.*

We expect this research direction to yield important, clinically relevant new results, but also to pose new questions for machine learning, data-mining and knowledge discovery.

Acknowledgments

The financial support of the European Commission (Project INFOBIOMED NoE, FP6/2004/IST-507585, and Project ACGT,FP6/2004/IST-026996) is gratefully acknowledged.

References

1. Sander, C.: Genomic Medicine and the Future of Health Care. Science **287**(5460) (2000) 1977–1978

2. Martin-Sanchez, F., et al.: Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. Journal of Biomedical Informatics **37**(1) (2004) 30–42
3. Foster, I., Kesselman, C., eds.: The Grid: Blueprint for a New Computing Infrastructure. 2nd edn. Morgan Kaufmann (2004)
4. Stankovski, V., May, M., Franke, J., Schuster, A., McCourt, D., Dubitzky, W.: A service-centric perspective for data mining in complex problem solving environments. In: Proc. Int. Conf. on Parallel and Distributed Processing Techniques and Applications (PDPTA'04). Volume II., Las Vegas, USA (2004) 780–787
5. Parks, M.R., Disis, M.L.: Conflicts of interest in translational research. Journal of Translational Medicine **2**(28) (2004) 1–4
6. Witten, I., Frank, E.: Data Mining – Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann (2000)
7. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2005) ISBN 3-900051-07-0.
8. Tsiknakis, M., Kafetzopoulos, D., Potamias, G., Analyti, A., Marias, K., Manganas, A.: Building a European Biomedical Grid on Cancer: The ACGT Integrated Project. Stud Health Technol Inform. **120** (2006) 247–258
9. Potamias, G., Tsiknakis, M., Papoutsidis, V., Kanterakis, A., Marias, K., Kafetzopoulos, D.: Advancing Clinico-Genomic Research Trials via Integrated Knowledge Discovery Operations. In: MIE2006, (poster presentation). (2006)
10. Potamias, G., Koumakis, L., Moustakis, V.: Mining XML Clinical Data: The HealthObs System. Ingenierie des systems d'information, special session: Recherche, extraction et exploration d'information **10**(1) (2004) 59–79
11. Potamias, G., Koumakis, L., Moustakis, V.: Gene Selection via Discretized Gene-Expression Profiles and Greedy Feature-Elimination. LNAI **3025** (2004) 256–266
12. Eisen, M., Spellman, P., Botstein, D., Brown, P.: Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. **96** (1999) 14863–14867
13. Alizadeh, A., et al.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature **403** (2000) 503–511
14. Golub, T., et al.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science **286** (1999) 531–537
15. Alon, U., et al.: Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. Proc. Natl. Acad. Sci. **96** (1999) 6745–6750
16. Gupta, S., Rao, S., Bhatnagar, V.: K-means Clustering Algorithm for Categorical Attributes. LNCS **1676** (1999) 203–208
17. San, O.M., Huynh, V., Nakamori, Y.: An alternative extension of the k-means algorithm for clustering categorical data. Int. J. Appl. Math. Comput. Sci. **14**(2) (2004) 241–247
18. Kanterakis, A., Potamias, G.: Supporting Clinico-Genomic Knowledge Discovery: A Multi-Strategy Data Mining Process. LNAI **3955** (2006) 520–524
19. Katehakis, D., Sfakianaki, S., Tsiknakis, M., Orphanoudakis, S.: An Infrastructure for Integrated Electronic Health Record Services: The Role of XML. Journal of Medical Internet Research **3**(1) (2001) E7
20. van't Veer, L., et al.: Gene expression profiling predicts clinical outcome of breast cancer. Nature **415** (2002) 530–536