19th IEEE International Conference on Tools with Artificial Intelligence

Knowledge Discovery Scientific Workflows in Clinico-Genomics

George Potamias[1], Lefteris Koumakis[1], Alexandros Kanterakis[1,3], Stelios Sfakianakis[1], Anastasia Analyti[1], Vassilis Moustakis[1,3], Dimitris Kafetzopoulos[2], Stefan Rueping[4], Manolis Tsiknakis[1]
[1]*Institute of Computer Science (ICS),* [2]*Institute of Molecular Biology & Biotechnology (IMBB), Foundation for Research & Technology – Hellas (FORTH), Crete, Greece*
{potamias,kounakis,kantale,ssfak,analyti,moustaki,tsiknaki}@ics.forth.gr, kafetzo@imbb.forth.gr
[3]*Technical University of Crete, Department of Production Engineering & Management, Greece*
[4]*Fraunhofer IAIS, Schloss Birlinghoven, 53754 St. Augustin, Germany*
stefan.rueping@iais.fraunhofer.de

Abstract

With the completion of the human genome and the entrance into the post-genomic era, translational research rises as a major need. In this paper, we present a Knowledge Discovery workflow (KDw) and its utilization in the context of clinico-genomic trials. KDw aims towards the discovery of 'evidential' correlations between patients' genomic and clinical profiles. Application of KDw on a real-world clinico-genomic (breast cancer) study demonstrates the reliability, efficacy, and efficiency of the approach.

1. Introduction

Biomedical research has already crossed the gate of laboratory bench, moving next to the patient bedside. The shift has created a new branch of research, termed *translational* research that catalyzes the clinical environment. The vision is to combat major diseases, such as cancer, on an *individualized* diagnostic, prognostic, and treatment manner. This requires not only an understanding of the genetic basis of the disease but also the correlation of genomic data with knowledge normally processed in the clinical setting [5]. Coupling the knowledge gained from genomics and from clinical practice is of crucial importance and presents a major challenge for on-going and future clinico-genomic trials [13]. Such *evidential* knowledge will augment health care professionals' decision-making capabilities in an attempt to support *evidence-based* medicine.

However, post-genomics advances have resulted in an explosion of information with the consequent shift to investigation methodology: from hypothesis-driven to data-driven with a focus on the search of biologically relevant patterns. In this dynamic and data-rich environment, the process of *biomedical knowledge discovery* calls for the seamless and flexible integration of both clinical and genomic disciplines, where clinicians, biomedical researchers, and data mining researchers exploit data from several diverse sources.

Towards meeting the aforementioned challenges, we elaborate on a technology-driven integrated clinico-genomic knowledge discovery process and its utilization in the context of clinical trials. The process aims towards the discovery of 'evidential' correlations between patients' genomic and clinical profiles, and it is realized by the device of a scientific *Knowledge Discovery workflow* (KDw). KDw is developed and implemented in the context of an operational integrated clinico-genomics environment that encompasses components and services for: (i) the seamless mediation between distributed and heterogeneous clinical and genomic data sources, and (ii) the *harmonization* of interoperable data mining operations [6].

2. Data Management and Mediation

KDw utilizes two *Clinical Information Systems* (CISs): (a) an *Onco-Surgery* information system that manages information related to BRCA patient identification and demographic information, medical history, patient risk factors, family history of malignancy, clinical examinations and findings, results of laboratory exams (mammography, biochemical exams, etc.), pre- and post-surgical treatment, as well as therapy effectiveness; and (b) a *Histo-Pathology* information system that manages information related to patients samples' histopathological evaluation and TNM staging (Tumor size, lymph Node involvement, and Metastatic spread).

1082-3409/07 $25.00 © 2007 IEEE
DOI 10.1109/ICTAI.2007.32

Moreover, KDw utilizes a *Genomic Information System* (GenIS) to store and manage microarray/gene-expression data. The system is based on the BioArray Software Environment BASE [10]. BASE is a comprehensive web-based database server that stores and manages the massive amounts of data, generated by microarray experiments.. We have extended and enhanced BASE in order to provide more advanced functionality (e.g., improved annotation of results; qualification of experiments; and new reporter/gene annotations to Gene Ontology (GO) (http://www.geneontology.org/).

Horizontal integration of the engaged heterogeneous information systems (CISs and GenIS) is achieved through a Web-based data mediation application - the *Mediator* (http://www.ics.forth.gr/~analyti/PrognoChip/isl_site/) [1]. The (authorized) biomedical investigator can form clinico-genomic queries through the web-based graphical user interface of the Mediator. The query specifies criteria for selecting patients/samples (and their corresponding clinical and genomic information) that have a specific clinico-histopathology profile and participate in microarray experiments of specified quality and characteristics. Clinical and genomic sub-query results are joined, based on the reference IDs of the retrieved patient samples, and an output XML file of predetermined schema is created. Since the set of selected reporters can be very large, reporter/gene annotations are stored in a separate tab-delimited file. Similarly, the retrieved gene expression data for each sample are stored in corresponding tab-delimited files. These files compose the basic input to the data mining operations (Section 3).

The Mediatior as Web-service. The presented mediator infrastructure is deployed as a Web-based application. For the realization of KDw, we built a web service that receives all the user-interface actions and invokes the execution engine of the mediator. The mediator executes the queries and stores the final results (the aforementioned output files) in the web server. The results can be retrieved through a URL that is in the disposal of the KDw components and data-analysis services.

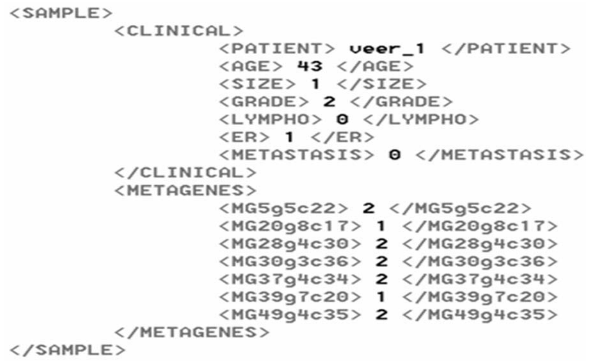
3. Data Mining Operations and Services

Characterization and classification of a disease and prediction of respective patients' clinical outcome may be achieved via reference either to solely standard clinical patient profiles/phenotypes (CPPs) or, to solely genomic/gene-expression profiles/phenotypes (GPPs). Starting from observable clinical disease states, the quest targets the identification of respective *molecular signatures* or *gene markers* able to discriminate between different disease states. Based on the central-dogma of molecular biology, CPPs could be fully 'shaped' and causally determined by respective GPPs. In this setting, the quest is forwarded towards the following task: "*which clinical phenotypes relate and (how they do so) with specific gene-expression phenotypes*?" Such a discovery-driven scenario falls into the individualized medicine context. GPPs may be utilized to 'screen' respective CPPs and refine the clinical decision making process (leading to the identification of specific patient groups that are more suitable for specific clinical treatment and follow-up procedures). The entire endeavor calls for the identification of abductive inferential rules that engage both CPPs and GPPs. In order to ease the discovery of such evidential associations and rules we elaborate on a layered process, realized by the smooth integration of two data-mining operations: clustering and association rules mining.

3.1 Indicative Gene-Clusters: the Metagenes

With the utilization of a clustering operation, we aim to induce indicative clusters of genes that meet a special characteristic: all of its genes exhibit a 'strong' gene-expression profile for all of its linked samples, i.e., exhibit solely 'high'/'UP'-regulated or solely 'low'/'DOWN'-regulated expression levels. We refer to these clusters of genes as *Metagenes*. Motivated by results in (i) the identification of co-regulated groups or clusters of genes, (ii) the discriminatory decomposition of genes, and (iii) the reduction of the dimensionality and complexity of gene-expression data [14,16,18], we utilize a *categorical k-means* clustering algorithm, named *discr_k-means*, that primarily identifies clusters of co-regulated binary-valued genes [2,12]. In order to overcome the error-prone variance of gene-expression levels, gene-expression values are *discretized* (following a data pre-processing discretization step) into two nominal values: 'low' and 'high'.

Clustering convergence provides a set of *rank-ordered* clusters; ordering connotes to cluster strength. Applying a filtering operation, we keep just those clusters (the Metagenes) for which all of its genes exhibit, in an adequate number of samples, 'strong' gene-expression profiles. We say that a sample has a '*strong*' gene-expression profile for a specific cluster of genes, if it exhibits dominantly 'high' or 'low' gene-expression levels. We are interested in such '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. That is why we decide to discretize the continuous gene expression levels into two nominal values and get the binary-valued gene-expression data transform. The genes of a metagene, accompanied by the 'strong' samples of the cluster, may be interpreted as a combined/*hybrid clinico-genomic feature*, linking patient cases and their genomic (gene-expression) profiles. For example, Figure 1 presents an XML-formatted hybrid clinico-genomic patient case: MG20g8c17= DOWN (value code '1' represents the DOWN-regulated status of a gene) denotes a metagene with id=20 ('MG20') that includes 8 genes ('g8'), covers 17 cases ('c17'), and for all 17 cases all the respective genes exhibit a DOWN value.



```
<SAMPLE>
        <CLINICAL>
                <PATIENT> veer_1 </PATIENT>
                <AGE> 43 </AGE>
                <SIZE> 1 </SIZE>
                <GRADE> 2 </GRADE>
                <LYMPHO> 0 </LYMPHO>
                <ER> 1 </ER>
                <METASTASIS> 0 </METASTASIS>
        </CLINICAL>
        <METAGENES>
                <MG5g5c22> 2 </MG5g5c22>
                <MG20g8c17> 1 </MG20g8c17>
                <MG28g4c30> 2 </MG28g4c30>
                <MG30g3c36> 2 </MG30g3c36>
                <MG37g4c34> 2 </MG37g4c34>
                <MG39g7c20> 1 </MG39g7c20>
                <MG49g4c35> 2 </MG49g4c35>
        </METAGENES>
</SAMPLE>
```

Figure 1. An XML-formatted 'hybrid' clinico-genomic (BRCA) patient case description

3.2. Discovery of Clinico-Genomic Associations

The task now is to uncover potential '*causal*' relations that hold between such genomic and clinical profiles. We handle this task with an association rules mining approach. *Association rules mining* (ARM) [8] is among the most advanced and interesting methods for finding interesting patterns and indicative trends in data. Given a set of transactions, the ARM problem is to discover the associations that have support and confidence values higher than the user specified minimum support and minimum confidence levels, respectively. For the implementation of ARM, we used *HealthObs* [7], which implements ARM operations on-top of XML documents. Central to the architecture is a single data-enriched XML document that contains query-specific data from distributed and heterogeneous data sources. ARM operations are performed on-top of such documents. The implemented ARM operations rely on the principles of the Apriori algorithm [8].

HealthObs as Web-service. The core functionality and operations of the originally stand-alone (java-based) HealthObs application were implemented as a Web-service, in order to serve KDw. The implementation takes as input four basic arguments (as an XML-formatted WSDL definition): the desired format of each rule, i.e., the features to be included in the 'IF' or 'THEN' part of the rules (a special characteristic offered by HealthObs); the minimum support and confidence of the rules, and a string that points to the data URL path. The definition of the input to HealthObs Web-service follows the BPEL4WS specifications (http://www.ibm.com/ developerworks/library/specification/ws-bpel/). The output of the HealthObs Web-service takes also a WSDL definition. The corresponding XML schema provides links to the path where the results (i.e., induced association rules) are stored, as well as to the path where the rules' visualization component of HealthObs is rest.

4. Editing and Enactment of KDw

Each workflow manages the execution of the various components that comprise a certain biomedical scenario. Possible components include database connections, data-mining and visualization tools, meta-data managing systems etc. KDw is realized by two basic components: the mediator and the data-mining (discr_*k*-means and HealthObs) web-services.

The *workflow editor* is an end-user component that provides the workflow authoring functionality by a user friendly GUI. Realization of KDw is based on the Taverna Workflow Editor [17]. It is essential to mention that the workflow execution is detached from the workflow authoring, i.e., the workflow is not executed on the user's desktop and the various by-products of the execution (results, breakpoints, provenance, etc.) are managed and persisted in an arbitrary workflow execution environment (i.e., Grid, Personal Computer, Web Server, etc.). Each workflow can be seen as a service with a single operation ('enact-me') that accepts the workflows parameters and returns the workflow results. Again, we rely on Taverna's Workflow Enactment service for the execution of KDw. KDw is deployed and runs on a Globus-based Grid infrastructure (http://www.globus. org/toolkit/).

5. KDw in Action

We applied and assessed the utility of KDw on a real-world, and widely utilized reference breast cancer (BRCA) study [11]. The study profiles the expression of ~24800 genes on 78 patients aiming towards the identification of highly discrimant molecular signatures (i.e., gene markers), able to distinguish: (i) between 'good' and 'bad' prognosis patients' cases, i.e., 'NO-Metastasis' and 'YES-Metastasis', respectively, in a follow-up period of five years; and (ii) between ER^+ vs. ER^- (i.e., estrogen receptor positive vs. estrogen receptor negative) patients. Here, we concentrate on both tasks with emphasis on the ER status, which provides a powerful predictive and prognostic BRCA marker [15].

Our aim was to reveal (potentially) interesting *clinico-genomic profiles* that combine patients' ER-statuses (i.e., their clinical profile) with indicative metagenes (i.e., their genomic profile). The target was to increase the confidence of prognostic clinico-genomic markers.

With the KDw Mediator component, we retrieved and filtered-out genes on the basis of the following criteria: 2-fold difference and p-value ≤ 0.01 on at least five patient cases (similar criteria were applied in the original reference study). A set of ~1000 genes was reserved. Then, the discr_k-means component was applied on this set of genes with the following parameterization: (a) discretization into two intervals – 'low'/'DOWN'-regulated and 'high'/'UP'-regulated gene-expression levels; (b) seeking for fifty metagenes - a high number of clusters is requested in order to adequately cover most of the cases and assign them to metagenes. A metagene is kept if: (i) all (100%) of its genes exhibit 'low' or 'high' expression levels, and (ii) it covers at least fifteen strong cases (~20% of the total cases).

A set of 22 metagenes were induced that cover 77 cases. With the KDw HealthObs component, we try to induce association rules that match the following format:

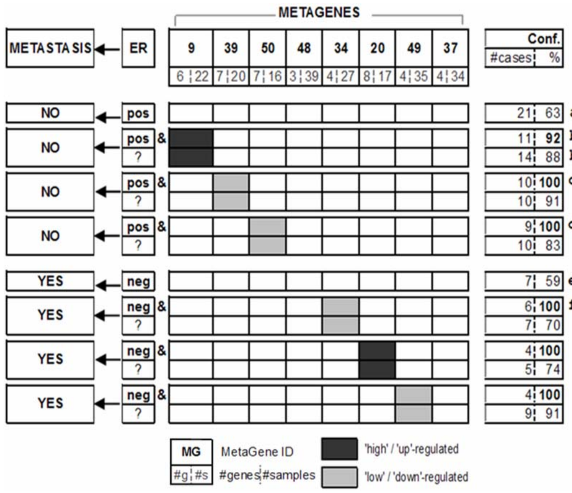
```
IF    ER-status  = pos/neg
    & Metagene_i = UP/DOWN
    & Metagene_j = UP/DOWN
      …
THEN  METASTASIS = NO/YES
```

With a minimum support and confidence of 5% and 50%, respectively, a set of 193 such rules were induced. As it is already mentioned, we are interested for highly-confident associations, when specific metagenes are included in the 'IF' part of the corresponding rules.

With this in mind and inspecting visually the rules, we were able to identify a set of prognostic and potentially interesting, rules (some of them are shown in Table 1).

The first rule (a) indicates that "*with a confidence of* 63%*, if* ER = pos *then* NO*-metastasis is expected within five years*" – the rule covers 21 cases. This rule could be contrasted with the second rule (b1) stating that "*with a confidence of* 92%*, if* ER = pos *and* Metagene_9 = UP (denoted with the black-bolted cell) *then* NO*-metastasis is expected in five years*" – the rule covers 11 cases. Even if less cases are covered (11 vs. 21), the drastic increase in confidence is profound.

Table 1. Indicative Associations/Prognostic Rules



			METAGENES								Conf.		
METASTASIS	ER		9	39	50	48	34	20	49	37	#cases	%	
			6¦22	7¦20	7¦16	3¦39	4¦27	8¦17	4¦35	4¦34			
NO	pos										21	63	a
NO	pos	&	■								11	**92**	b1
	?		■								14	88	b2
NO	pos	&		▒							10	**100**	c
	?			▒							10	91	
NO	pos	&			▒						9	**100**	d
	?				▒						10	83	
YES	neg										7	59	e
YES	neg	&					▒				6	**100**	f
	?						▒				7	70	
YES	neg	&						■			4	**100**	
	?							■			5	74	
YES	neg	&							▒		4	**100**	
	?								▒		9	91	

MG MetaGene ID
#g¦#s #genes¦#samples
■ 'high' / 'up'-regulated
▒ 'low' / 'down'-regulated

Moreover, when just 'Metagene_9 = UP' (third rule, b2) is used (i.e., no reference to ER status is made), our confidence for a good prognosis is decreased (88% vs. 92%). This is an indication for the need to combine both clinical (in this case histology) and genomic profiles in order to induce and devise reliable prognostic models. The situation becomes more drastic with the subsequent rules: rules c and d conclude a good prognosis with 100% confidence, when 'Metagene_39' or 'Metagene_50' is present in a 'DOWN'-regulated gene expression status ('DOWN' is denoted with a gray-bolted cell). Analogously, rule e is 59% confident for a bad prognosis, when just the ER-status is considered. Contrasting rule e with rule f, we note that: if 'Metagene_34' is also engaged in a 'DOWN'-regulated status then the confidence for a bad prognosis becomes 100%.

We also performed a biological assessment and validation of the discovered rules. For example, 'Metagene_39' (involved in rule c) includes seven genes, four of which could be found in public databanks, namely: EPHA8, COX7C, ECM1, MYH8 (note that all these genes exhibit a 'DOWN'-regulated profile in Metagene_39). Down-regulation of EPHA8 is important for invasiveness [9]; moreover it is found that its over-expression can increase metastatic potential [4]. Several studies have suggested that over-expression of COX-2 is associated with parameters of aggressive BRCA [3]. ECM1 is found to promote angiogenesis and its over-expression results to tumor growth and metastasis in BRCA cells [20]. Finally, it was recently found that two major myosin class-II isoforms (in which MYH8 is a member) are both expressed in metastatic BRCA cells [19].

Findings demonstrate the potential of the presented KDw approach in the course of clinico-genomic biomedical research. We believe that following the presented exploration methodology and the accompanying biological validation of results, KDw presents a flexible tool for the discovery of reliable clinico-genomic profiles and associations. The entire KDw (excluding the mediator component to retrieve data from the respective information systems) runs in the scale of few minutes.

6. Conclusions

With the use and customization of scientific workflow methodologies, enabled by Web-services technology, we devised and presented an integrated clinico-genomic knowledge discovery workflow (KDw). KDw relies on the smooth integration of clinico-genomic data mediation and data mining components. Overall, KDw composition is based on the customization of Web-services for its corresponding components, and the utilization of the Taverna workflow editing and enactment environment. KDw composes a flexible and adaptive tool that supports exploration and knowledge discovery activities in the context of post-genomic clinical trials. Application of the KDw on a real-world breast cancer study, followed by a careful exploration methodology, demonstrates the reliability and utility of the whole approach.

Acknowledgment. Work reported herein is partially supported via the ACGT Integrated Project (FP6-2005-IST-026996; http://www. eu-acgt.org), and the INFOBIOMED Network of Excellence (FP6-IST-507585; http://www. infobiomed.org).

References

[1] A. Analyti, *et al.*, “Integrating Clinical and Genomic Information Through the PrognoChip Mediator”, ISBMDA’06, *LNCS* 4345, 2006, pp. 250-261.

[2] A. Kanterakis, and G. Potamias, “Supporting Clinico-Genomic Knowledge Discovery: A Multi-Strategy Data Mining Process”, SETN’06, *LNCS* 3955, 2006, pp. 520-524.

[3] B. Arun, and P. Goss, “The role of COX-2 inhibition in breast cancer treatment and prevention”, *Semin Oncol.* 31(2 Suppl 7), 2004, pp. 22-29.

[4] D.P. Zelinski, *et al.*, “Estrogen and Myc negatively regulate expression of the EphA2 tyrosine kinase”, *J Cell Biochem.* 85(4), 2002, pp. 714-720.

[5] F. Martín-Sánchez, *et al.*, “Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care”, *Journal of Biomedical Informatics* 37(1), 2004, pp.30-42.

[6] G. Potamias, D. Kafetzopoulos, M. Tsiknakis, “Integrated Clinico-Genomics Environment: Design and Operational Specification”, *Journal for Quality of Life Research* 2(1), 2004, pp. 145-150.

[7] G. Potamias, L. Koumakis, and V. Moustakis, “Mining XML Clinical Data: the HealthObs System. *Ingénierie des Systèmes d'Information* 10(1), 2005, pp. 59-79.

[8] H. Mannila, H. Toivonen H., and A.I. Verkamo, ”Efficient algorithms for discovering association rules”, *AAAI Workshop on Knowledge Discovery in Databases* (KDD-94), Seattle, Washington, 1994, pp. 181-192.

[9] H. Surawska, P.C. Ma, and R. Salgia, “The role of ephrins and Eph receptors in cancer”, *Cytokine & Growth Factor Reviews* 15(6), 2004, pp. 419-433.

[10] L.H. Saal, *et al.*, “BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data”, *Genome Biology* 3(8): software0003.1–0003.6, 2002.

[11] L.J. van ’t Veer, *et al.*, “Gene expression profiling predicts clinical outcome of breast cancer”, *Nature* 415, 2002, pp. 530-536.

[12] M. May, G. Potamias, and S. Rüping, “Grid-based Knowledge Discovery in Clinico-Genomic Data”, *Lecture Notes in Bioinformatics* 4345, 2006, pp. 219-230.

[13] M. Tsiknakis, D. Kafetzopoulos, G. Potamias, A. Analyti, K. Marias, and A. Manganas, “Building a European biomedical grid on cancer: the ACGT Integrated Project”, *Stud Health Technol Inform.*, 120, 2006, pp. 247-258.

[14] M.B. Eisen, D.T. Spellman, P.O. Brown, and D. Botstein, “Cluster analysis and display of genome-wide expression patterns”, *Proc. Natl. Acad. Sci.* 96, 1999, pp. 14863-14867.

[15] S. Sommer, and S.A. Fuqua, “Estrogen receptor and breast cancer”, *Semin Cancer Biol.* 11(5), 2001, pp. 339-352.

[16] T. Golub, *et al.*, “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring”, *Science* 286, 1999, pp. 531-537.

[17] T. Oinn *et al.*, **“**Taverna: a tool for the composition and enactment of bioinformatics workflows”, *Bioinformatics* 20(17), 2004, pp. 3045-3054.

[18] U. Alon, 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, pp. 6745-6750.

[19] V. Betapudi, L.S. Licate, and T.T. Egelhoff, “Distinct Roles of Nonmuscle Myosin II Isoforms in the Regulation of MDA-MB-231 Breast Cancer Cell Spreading and Migration”, *Cancer Research* 66, 2006, pp. 4725-4733.

[20] Z. Han, *et al.*, “Extracellular matrix protein 1 (ECM1) has angiogenic properties and is expressed by breast tumor cells”, *FASEB* 15, 2001, pp. 988–994.