PROCEEDINGS

George Potamias
Vassilis Moustakis
(eds.)

Heraklion, Crete, Greece
April 2009
BMIINT Editorial

SCENE. Entering the post-genomics epoch a new challenging mission is posted: to bring innovative biomedical research findings directly to the clinic and the bedside. The mission could be accomplished by intersecting the clinical, biological and information sciences. Endeavour is inspired by the needs raised by genomic and personalised medicine; it targets the in-silico biology domain; and, it is enabled by the transition to interdisciplinary principles and orientation. Transition is guided by the ‘-omics’ revolution as realised by advances in transcriptomics, proteomics, metabolomics, physiomics and nanomedicine. The ‘-omics’ levels extends traditional clinical data models and medical decision making in two sides: on the one side to include genotypes, and on the other to include (when appropriate) ‘-omics’ findings - the phenotypes. Yet integration is not easy simply because at both ends (genotype, phenotype) the amount of available data is immense, and complexity of processes is high. At the up-stream (or research) level scientists may be interested in the investigation between genotype-phenotype information to form associations and patterns of disease and susceptibility indices. At the middle-stream (industrial level) technology and service providers are interested in embedding research advances into concrete products and services via intelligent devices. Device impact is further enhanced by advances in the ‘nano’ field, as expressed by nanomaterials, nanomedicine and nanoinformatics. At the down-stream (clinical theatre) level healthcare professionals (and patients as well) look forward to apply new technology on the vision of continuously improving social welfare decision making.

Actors & Director. The primary item in the respective multidisciplinary R&D agenda is translational research with Biomedical Informatics (BMI) as the driver, and Artificial Intelligence (AI) called to provide the needed analytical and decision-making machinery. BMIINT liaisons the clinical, biology, core- and Bioinformatics fields, and provides a forum for the presentation of advances at the conjunction of BMI and AI components and procedures. It takes an interdisciplinary perspective and focuses on theory, methods, techniques, tools, systems and services, which support integration, management and intelligent analysis of respective bio-related information and data sources. Emphasis is placed on the repositioning of methods and techniques from other domains of application into the BMIINT frontier.

Plot(s). BMIINT includes nine papers, which report on R&D work with broad coverage of BMIINT context. Article by Martin-Sanchez and colleagues, reports the organizational aspects and scientific aspects in relation to the COMBIOMED: A Cooperative Thematic Research Network on COMPUTATIONAL BIOMEDicine in Spain. The network focuses on gene-disease associations, pharmainformatics and decision support systems at the point of care. Work reported by Bonsma and Vrijnsen; Exarchos, Goletsis and Fotiadis focus on heterogeneous data integration. Bonsma and Vrijnsen provide an enrichment of the OGSA-DAI web-services framework to access medical images and microarray data; work carried out in the context of Advanced Clinico-Genomic Trials in Cancer (ACGT) project (FP6-IST-2005-026996). Exarchos et al., integrate clinical, imaging and genomic data sources to induce reliable biomarkers for the progression of oral cancer – work relates to the NeoMark project (FP7-ICT-2007-224483). Related to the two aforementioned papers is the work presented by Sfakianakis and colleagues as well by Koumakis and colleagues. Sfakianakis and colleagues endeavour on the
daily work of clinicians and bio-statisticians, develop usability criteria and propose a front-end interface layer for a Grid based architecture able to support huge computational tasks. The work of Koumakis and colleagues takes a step backwards and captures scientific-workflow design and operation specifics with due regard to a clinical scenario that achieves the seamless integration of clinicogenetic heterogeneous data sources, and the discovery of indicative SNP-phenotype associations and predictive models. Both articles draw from the ACGT project experience while Koumakis and colleagues work relates also to the GEN2PHEN project (co-funded via the European Commission, Health theme, project number 200754). Remaining articles focus on more specific BMIINT subjects. Tsiliki and colleagues overview requirements and context of cross-platform integration; the subject has immense interest, given the multitude of microarray platforms. Blachon and colleagues as well as Kanterakis and colleagues bring into BMIINT a systems biology flavor. Blachon and colleagues work is on the Ewing sarcoma; using a comparative genomic hybridization array they present data collection and preprocessing procedures and then move on to a gene influence network to model discovery of links between gene copy number variations and expression level variations – work relates to the SITCON project (from the ANR BIOSYS-2006 program). Kanterakis and colleagues reports on a methodology that couples gene-regulatory networks and microarray gene-expression data to reveal and identify molecular paths that differentiate between different disease phenotypes (with targets to the Wilms tumor domain) – work also relates to the ACGT project. Finally, Vegoudakis and colleagues report on an interface that supports patterns matching over genomic and proteomic sequences on a Grid based system – work relates to the EGEE project, co-funded by the European Commission.

**Scientific Committee.** Organization of BMIINT was supported by an international Scientific Committee. Members of the committee are:

- **Anastasios Bezerianos**, University of Patras/Medical School, GR
- **Pierre-Alain Binz**, Swiss Institute of Bioinformatics & GENEBIO, CH
- **Anthony Brooks**, University of Leicester, UK
- **Anca Bucur**, Philips Research, NL
- **Francesca Buffa**, University of Oxford, UK
- **Alberto D’Onofrio**, European Institute of Oncology, IT
- **Dimitris Fotiadis**, University of Ioannina, GR
- **Pierre Grangeat**, Commissariat a l’Energie Atomique (CEA), FR
- **Artemis Hatzigeorgiou**, B.S.R.C. “Alexander Fleming”, GR
- **Dimitris Kafetzopoulos**, FORTH-IMBB, GR
- **Antonis Kakas**, University of Patras & RACTI, GR
- **Elpida Keravnou**, University of Cyprus, CY
- **Josipa Kern**, University of Zagreb, HR
- **Maria Klapa**, FORTH-ICEHT, GR
- **Marja Laine**, University of Amsterdam/ ACTA, NL
- **Aristidis Likas**, University of Ioannina, GR
- **Nikos Maglaveras**, Aristotle University of Thessaloniki, GR
- **Victor Maojo**, Universidad Politecnica de Madrid (UPM), ES
- **Fernando Martin-Sanchez**, Instituto de Salud Carlos III (ISCIII)/BIOTIC, ES
- **Luciano Milanesi**, National Research Council (CNR) / ITB, IT
- **Konstantina Nikita**, National technical University of Athens (NTUA), GR
- **Helen Parkinson**, EMBL / European Bioinformatics Institute (EBI), UK
- **George Patrinos**, Erasmus MC / Faculty of Medicine and Health Sciences, NL
- **Stefan Rueping**, Fraunhofer / IAIS, DE
Elias Siores, University of Bolton, UK
Srdjan Stankovic, University of Belgrad, YU
Ioannis Tollis, University of Crete & FORTH / Institute of Computer Science (ICS), GR
Oswaldo Trelles, University of Malaga, ES
Ioannis Tsamardinos, University of Crete & FORTH-ICS, GR
Manolis Tsiknakis, FORTH / Institute of Computer Science (ICS), GR
Michael Zervakis, Technical University of Crete, GR

The Chairs of AIAI 2009/BMIINT

Dr. George Potamias, Institute of Computer Science, BMI Lab, FORTH – Heraklion, Greece

Prof. Vassilis Moustakis, Institute of Computer Science, BMI Lab, FORTH & Technical University of Crete, Department of Production Engineering and Management, Management Systems Laboratory, Chania, Greece

Crete, April 2009

ACKNOWLEDGEMENT: The Chairs want to thank the members of the Scientific Committee for their support. In addition, organization of BMIINT is partially supported by the ACTION-Grid EU project (FP7 ICT 224176), as well by the ACGT and GEN2PHEN projects (as mentioned in the preceding text).
## Contents

<table>
<thead>
<tr>
<th>Paper</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMBIOMED: A Cooperative Thematic Research Network on Computational Biomedicine in Spain</td>
<td>1</td>
</tr>
<tr>
<td>Fernando Martín-Sánchez, Victoria Lopez-Alonso, Isabel Hermosilla-Gimeno, Guillermo Lopez-Campos, and the COMBIOMED Network</td>
<td></td>
</tr>
<tr>
<td>Homogenising access to heterogeneous biomedical data sources</td>
<td>9</td>
</tr>
<tr>
<td>Erwin Bonsma and Jeroen Vrijnsen</td>
<td></td>
</tr>
<tr>
<td>Unification of heterogeneous data towards the prediction of oral cancer reoccurrence</td>
<td>21</td>
</tr>
<tr>
<td>Konstantinos P. Exarchos, Yorgos Goletsis, Dimitrios I. Fotiadis</td>
<td></td>
</tr>
<tr>
<td>Building a System for Advancing Clinico-Genomic Trials on Cancer</td>
<td>33</td>
</tr>
<tr>
<td>Stelios Sfakianakis, Norbert Graf, Alexander Hoppe, Stefan Raping, Dennis Wegener, Lefteris Koumakis, and George Zacharioudakis</td>
<td></td>
</tr>
<tr>
<td>Discovery of Genotype-to-Phenotype Associations: A Grid-enabled Scientific Workflow Setting</td>
<td>45</td>
</tr>
<tr>
<td>Lefteris Koumakis, Stelios Sfakianakis, Vassilis Moustakis, and George Potamias</td>
<td></td>
</tr>
<tr>
<td>Cross-platform integration of transcriptomics data</td>
<td>57</td>
</tr>
<tr>
<td>Georgia Tsiliki, Marina Ioannou, and Dimitris Kafetzopoulos</td>
<td></td>
</tr>
<tr>
<td>Method for relating inter-patient gene copy numbers variations with gene expression via gene influence networks</td>
<td>69</td>
</tr>
<tr>
<td>Sylvain Blachon, Gautier Stoll, Carito Gaziolowski, Andrei Zimovyev, Emanuel Barillot, Anne Siegel, and Ovidiu Radulescu</td>
<td></td>
</tr>
<tr>
<td>Revealing Disease Mechanisms via Coupling Molecular Pathways Scaffolds and Microarrays: A Study on the Wilm’s Tumor Disease</td>
<td>85</td>
</tr>
<tr>
<td>Alexandros Kanterakis, Vassilis Moustakis, Dimitris Kafetzopoulos, and George Potamias</td>
<td></td>
</tr>
<tr>
<td>A Simple Algorithm Implementation for Pattern-Matching with Bounded Gaps in Genomic and Proteomic Sequences, on the Grid EGEE Platform, using an intuitive User Interface</td>
<td>97</td>
</tr>
<tr>
<td>K.I. Veggoudakis, K.G. Margaritis, and N. Maglaveras</td>
<td></td>
</tr>
</tbody>
</table>
COMBIOMED: A Cooperative Thematic Research Network on COMputational BIOMEDicine in Spain

Fernando Martin-Sanchez, Victoria Lopez-Alonso, Isabel Hermosilla-Gimeno, Guillermo Lopez-Campos, and the COMBIOMED Network.

Medical Bioinformatics Department, Institute of Health "Carlos III", Madrid, Spain
fms@isciii.es

Abstract

The Cooperative Thematic Research Network on Computational Biomedicine COMBIOMED was approved in the last call for Thematic Networks in Health Research within the Spanish National Plan for Scientific Research, Development and Technological Innovation and it is funded for the period 2008-2011. The COMBIOMED Network is currently addressing various aspects that range from basic to applied research in science for the development of methods and tools to solve problems in biomedical science in the context of personalized medicine. This paper describes and analyses the organizational aspects and scientific areas in which this network has been focused (gene-disease association, pharmainformatics and decision support systems at the point of care). At the same time, COMBIOMED aims to play a central role in the education of researchers and in the training of health professionals in techniques for the processing of biomedical information.

1. Introduction

The COMBIOMED Network continues the work initiated by INBIOMED, the Cooperative Thematic Research Network on Biomedical Informatics (2004-2007) that developed a platform for the storage, integration and analysis of clinical, genetic, and epidemiological data and images focused on the investigation of complex diseases [1]. Computational biomedicine represents the interface between biomedical and computer sciences. It provides an inclusive environment for a better understanding of the biological processes that take place in each of the levels of organization of living organisms and the intricate network interactions between...
them. One of the objectives of COMBIOMED is to establish contacts and to collaborate with the most relevant international initiatives in the field, such as the National Centers for Biomedical Computing [2], or the Biomedical Informatics Cores of the Clinical Science and Translational Awards [3] funded by the National Institutes of Health (NIH).

2. COMBIOMED: description and organization

Several of the 12 groups participating in COMBIOMED have previously participated and led some of the European Networks of Excellence (NoE) on biomedical informatics, bioinformatics and systems biology, such as INFOBIOMED [4] and BIOSAPIENS [5]. Previous experience in Spanish initiatives (INBIOMED [6], INB [7]) has been crucial to the creation and development of those European initiatives.

The design of the network, shown in Figure 1, consists of the following levels:

- Coordination and management.
- Computational aspects, which serve as instrumental support to the network including software and middleware, hardware, GRID, algorithms, and programming.
- Coordination of the work carried out by the research groups. This level addresses aspects such as data and text mining, clinical decision-making, electronic health records, image processing, disease simulation, and biomedical ontologies that help manage and integrate chemical, genetic, environmental, clinical and imaging data.
- Horizontal activities affecting all groups and lines of work. Particular attention is paid to the connection of the network with the scientific community and society (integrated Knowledge Management, education and training, communication, dissemination, Quality and Safety) (Figure 2).

3. COMBIOMED: scientific areas

COMBIOMED focuses on three scientific research areas: gene-disease association (Disease-omics), pharmainformatics and decision support systems at the point of care (Info-POC).
3.1 Gene disease association (Disease-omics)

The study of the molecular causes of disease and individual genetic variations allows deepening into personalized medicine [8] by developing safer and more efficient preventive, diagnostic and therapeutic solutions. The scientific community needs more advanced computational resources (functional analysis of genes and proteins in the context of genetic variability, alternative splicing...) [9], access to specific comparative genomic information (genomic data visualization) and prediction of the effects of individual mutations (SNPs) in the pathways and macromolecular complexes with the consequent implications in the associated diseases.

![Diagram](image)

**Figure 1.** Graphical representation of the structure of the COMBIOMED Network.

COMBIOMED works on these computational challenges in genotype-phenotype association and genomic epidemiology studies, to advance the understanding and modeling of the influence of environmental and genetic factors in the development of diseases. The network is using modules already developed by the National Institute of Bioinformatics (INB) to connect new methods that will be made available as Web services. This will help develop specific solutions for the analysis of genomic and clinical data.

The network is also developing systems that facilitate access to textual information about gene-disease relationships using automated information extrac-
tion methods and natural language processing with specific applications to problems of biomedical importance [10].

3.2 Pharma-informatics

The discovery and development of drugs is an area of great significance for human health, and at the same time it is an area of great socio-economic importance because it gives its “raison d’être” to an industry which business is highly knowledge-intensive. Biomedical research in general and the R & D of drugs in particular, generate enormous amounts of data that require sophisticated computational tools for their management and analysis in order to extract the knowledge they need. This is one of the main reasons for the emergence of a new field of scientific activity that includes disciplines such as Computational Biology, and Biomedical Informatics.

Pharmaceutical research labs were pioneers in identifying the need and usefulness of computational approaches for the management and exploitation of the data generated in pre-clinical and clinical research. They are aware that certain
computational methods and their associated software can perform simulations and predictions that save time and investment in the development of drugs [11]. Computational approaches in systems biology are facilitating the management, visualization and development of predictive and descriptive mathematical models on interaction networks between biomolecular entities. This information is generated in the experimental laboratory largely based on the use of microarrays technologies [12].

Virtual screening and computer simulation techniques are very useful for the selection and testing of compounds to be considered in the initial stages of the design of a new drug. Moreover, the pharmacological and toxicological knowledge accumulated on the different groups of compounds allows for the development of quantitative models that can be used to perform *in-silico* prediction studies of the pharmacological and toxicological behavior of compounds not synthesized or tested.

Information technology also plays an important role in areas such as the management and exploitation of data from clinical trials. In addition, physiological advanced simulation techniques may allow the study of the behavior of organs of different individuals when exposed to drugs with different properties.

In coordination with the INB and the Spanish Technological Platform of Innovative Medicines [13], the COMBIOMED Network is developing technological solutions to facilitate the advancement of biomedical knowledge management geared towards the development of pharmaceutical R & D in all its stages.

### 3.3 Decision support systems at the point of care (INFO-POC)

In recent decades medical practice has sought greater integration of scientific knowledge in its routine. The tremendous growth of scientific knowledge and technological innovation requires the development of solutions that allow the use of a large amount of information in the clinical decision-making process. Within this context, Computational Biomedicine promotes the combination of techniques such as Medical Informatics (MI), bioinformatics (BI) and computing in the development of new methods and standards for clinical and biomolecular data integration and analysis [14]. At the same time, they facilitate a new approach that has as its overall objective to create a new integrated research framework for the development of diagnostic methods within the context of genomic medicine in the so-called "point of care".

The COMBIOMED network proposes the common research line of INFO-POC to carry out computational developments to represent and analyze clinical and biomedical knowledge at the point of patient care (POC). The collaboration between the diverse groups of the COMBIOMED network makes possible a continuous exchange of information and tools.
The network will support decision-making processes in a context of miniaturization of diagnostic systems and accessibility to information about molecular causes of diseases. This context is in line with recent trends on Convergent Technologies NBIC (Nano, Bio, Info and Cogno) with the objective of contributing to the development of a line of intelligent and miniaturized systems to be used at the point of care.

The availability and applicability of new technologies at the point of care could be a key incentive for translational research which may also imply a reduction in the time devoted to decision-making.

The DNA microarray technology and the Bioinformatics tools that allow microarray data storage, management and analysis have enabled the development of diagnostic tests for complex diseases [15]. In addition to the biomolecular results obtained through these miniaturized point-of-care test systems there exists the requirement of placing molecular data (i.e. mutations in a gene, sequences of DNA, proteins...) in context, through the recovery of relevant information from reference databases (in silico), and its interpretation by implementing systems to support the diagnosis process (in info). The enormous complexity of cellular processes (metabolism, signal transduction, gene expression, and so on) needs the de-
velopment of new computational models and simulations to understand their behavior overall. The recent boost of systems biology and computational cell biology reflects this fact. The design of new computer-based methods in the Semantic Web for data recovery can contribute to the representation and computational analysis of biological knowledge at the POC. The knowledge generated will be integrated into computerized protocols for the diagnosis, treatment and management of patients (Figure 3).

The combination of bioinformatics and biomedical computing tools will facilitate the development of diagnostic models, supported by new standards. These tools need to be linked by using standard medical terminologies and coding with clear semantics to facilitate the effective implementation within clinical information systems.

Conclusions

The creation of the COMBIOMED Network represents a national and international reference in biomedical computing, which aims to provide solutions to the computational challenges posed by basic and translational research, and clinical practice in the context of the new personalized medicine. The most relevant research groups in Spain are cooperating to develop methods, systems, applications and pilot projects and to yield educational recommendations to promote biomedical computing research in the next years.

More specifically, computational developments within the COMBIOMED Network allow advancing in the representation and analysis of clinical and biomolecular knowledge, and the joint research will enable the new generation of miniaturized systems to support decision making with obvious clinical applications in health at the point of care.

Acknowledgments


References


Homogenising access to heterogeneous biomedical data sources

Erwin Bonsma and Jeroen Vrijnsen

Philips Research, High Tech Campus 37, 5656 AE Eindhoven, The Netherlands

Abstract This paper reports our experiences of developing data access services in the context of the ACGT project. The paper documents two aspects of the work that we carried out. First the focus is on the problem of how to best provide a syntactically homogeneous data access interface for a set of heterogeneous data sources. We describe related work, outline the approach we have taken, and report our findings. The second part of this paper documents integration issues that we encountered when realizing the data access services. Choices with regards to realization have significant impact on the time and effort that is needed to develop and maintain the services and our experiences may provide useful guidance to others wanting to develop similar functionality.

Introduction

The work reported here has been carried out in the context of the ACGT (Advancing Clinico-genomic Trials on Cancer) project. The aim of ACGT is to develop open-source, semantic and grid-based technologies in support of post-genomic clinical trials in cancer research. One of the main challenges in carrying out post-genomic research is to efficiently manage and retrieve all relevant data. Carrying out a post-genomic clinical trial involves the collection and storage of a wide variety of data, including: clinical data collected on Case Report Forms (e.g. symptoms, histology, administered treatment, treatment response), imaging data (e.g. X-Ray, CT, MR, Ultrasound), and genomic data (e.g. microarray data). Next to that there are many public biomedical databases that are relevant. These store information about gene and protein sequences, pathways, genomic variation, microarray experiments, medical literature, tumour antigens, protein domains, metabolites, etc. Biomedical researchers currently have to use many different tools.

1 ACGT is an Integrated Project funded in the 6th Framework Program of the European Commission under the Action Line “Integrated biomedical information for better health”
and web interfaces to find and extract the data that is relevant to their clinical research. Providing seamless and integrated access to clinical, genetic and image databases would therefore greatly facilitate post-genomic research.

In order to provide seamless access to a heterogeneous set of databases syntactic and semantic integration needs to take place. Syntactic data integration handles differences in the formats and mechanisms of data access, whereas semantic integration deals with the meaning of information; it must handle the fact that information can be represented in different ways, using different terms and identifiers.

With regards to syntactic heterogeneities, the main areas where databases differ are:

- access protocols, e.g. SOAP/HTTP, DICOM, JDBC,
- data formats, e.g. different formatting of date values,
- message formats, e.g. XML, HTML, protocol-specific, and
- query mechanisms, e.g. SQL, literal matching, keyword-based search, or protocol-specific.

An example of a query mechanism specific to the biomedical domain is BLAST [2], which is used by sequence databases. Matching is approximate and parameters can be specified controlling the accuracy and speed of matching. A completely different query mechanism is needed to access medical image data, which is standardised using the DICOM protocol [3]. DICOM does not allow complex queries, as it does not intend to provide a generalized database query mechanism [4]. The baseline query functionality is very basic, and the optional extended query functionality is still limited and eccentric.

Semantic integration in ACGT is handled using Query Translation, carried out by a semantic mediator that uses a Local as View approach. It accepts queries expressed in the ACGT Master Ontology, divides them in sub-queries, and translates each to the ontology used by the underlying database. The remainder of this paper focusses on the syntactic integration of data sources. For details about the semantic integration approach, please refer to [5].

Related work

Syntactically homogeneous access to distributed data sources is typically provided by way of wrappers [6, 7, 8, 9]. One of the main challenges in building wrappers is the variation in the query functionality of the underlying data sources [10]. Data sources may not only use different data models and syntactically different query mechanisms, but their query capabilities can differ as well. This makes it difficult to support a common query language, an essential step towards syntactic homogeneity. There are two extreme approaches [7]. A highly expressive common query language can be chosen. This, however, makes it difficult to implement wrappers for sources with primitive query capabilities. Furthermore, if the wrappers are used by a mediator, it means that query decomposition, subquery scheduling and
result composition may be done by both; the mediator must be able to decompose queries across multiple data sources and a wrapper for a data source must be able to decompose a complex query into simpler ones that the data source can handle. This means duplication of implementation effort but also leads to overall sub-optimal query execution performance. On the other hand, if a very basic common query language is chosen, significant and unnecessary performance penalties are introduced as the capabilities of the underlying data sources are not effectively used.

As neither approach is ideal, an intermediate solution is proposed in [7]. A powerful common query language is chosen, but wrappers may choose to only support a subset of the queries, based on the capabilities of the underlying data source. Each wrapper describes the queries it supports using the Relational Query Description Language (RQDL) developed for this purpose. An RQDL specification consists of a set of query templates that represent parameterized queries that are supported. RQDL uses a context-free grammar to describe arbitrarily large sets of templates. Templates can be schema-independent as well as schema-dependent. Benefits of this approach are that wrappers can provide and expose query functionality that better corresponds to that of the underlying data source. A drawback is the increased complexity associated with interpreting and reasoning about the query capabilities of each source, but feasibility is demonstrated by the Capabilities-Based Rewriter, described in the same paper, that uses the wrappers and produces query execution plans in reasonable time.

A more recent example, applied in practice to life sciences data, is given by DiscoveryLink [8], a database middleware system for extracting data from multiple sources in response to a single query. The system consists of two parts: a wrapper architecture, and a query optimizer. SQL is used as the common query language for the wrappers, but wrappers may only support a subset of SQL. In the simplest case, a wrapper retrieves a projection over all rows in a given table. Wrappers can, however, also indicate that they support filtering conditions, or joins, and if so, how many. The paper proposes to involve wrappers in the query optimization process. Wrappers are asked for estimates on the query execution time and expected size of the result set for different sub queries. The query optimizer will use this information when deciding how to decompose the query. It requires efficient communication between the query optimizer and the wrapper, which is made possible because wrappers are shared-libraries, co-located with the query optimizer.

EDUTELLA [11] uses an RDF query language that has various language levels, with increasing functionality. The basic level supports RDF graph matching, the level above that adds disjunction, and the use of recursion in queries is added at even higher levels. Support for aggregation is an optional feature, orthogonal to these levels. Wrappers can support the level of the query language that best fits the query capabilities of their data source.

It is generally recognized that writing wrappers requires significant programming effort, and as a result significant research efforts have been devoted to
automating parts of this (see e.g. [6], [9]). In general, automation is focused on a subset of the different data sources, e.g. sources with a web interface [12].

Approach

We identified the following functional requirements for the data access services. Firstly, they should provide a uniform data access interface. This includes uniformity of transport protocol, message syntax, query language, and data format. Secondly, they should export the structure of the database, using a common data model, together with possible query limitations of the data source. Clients of the web service require this information for constructing queries. Thirdly, they should enforce the data source access policy, and audit access to data sources. For post-genomic clinical trial data, there exist strict legal and ethical requirements that need to be adhered to.

A common query language is needed to achieve a uniform interface. It needs to meet various requirements. Firstly, it must be sufficiently expressive; it should support the types of queries that clinicians and biomedical researchers want to carry out. Secondly, it must be attainable, with acceptable effort, to map the query language to those used by the various data sources that need to be accessed. Thirdly, it must be convenient to use the query language for semantic mediation, the next step of the data integration process. Fourthly, it should be a community accepted standard. This ensures that there are sufficient support tools available, such as parsing and query engines, and also increases the possibilities for our approach to be eventually widely adopted. We have chosen SPARQL [13] as the query language, as it satisfies all these requirements.

Web Services have been chosen as the common interface technology within ACGT, as this technology suits the distributed nature of the project with respect to the data, computing resources, and development teams. For the data access services we decided additionally to use OGSA-DAI, a Web Services framework for data access [14]. It uses an activity framework that enables flexible service invocation, and re-use of common data access functionality. The results of queries will be returned using the SPARQL Query Results XML Format [15], which is the natural choice given the web services context and the use of SPARQL.

To meet the second requirement each data access service exports its schema using RDF Schema [16]. This is the standard way to describe RDF data sources, which is how the data sources appear given that SPARQL is used.

Access to each data source is controlled by integrating the data access service into the ACGT security infrastructure. Authentication is credential-based and delegation of credentials between services is supported. Authorization is controlled centrally and authorization decisions are, amongst others, based on membership to virtual organizations, which can be created as required.
Implementation

We have implemented data access services for three data source types: relational databases, medical image databases, and microarray databases. These databases have been chosen after careful review of requirements; they are considered the most important in the context of post-genomic clinical trials given the data-mining scenarios that were identified during the requirements-gathering process.

Figure 1 shows the data access services in the context of the data analysis architecture. The workflow enactor carries out data-mining workflows. It uses the semantic mediator for retrieving data. The latter accepts queries expressed in the ACGT Master Ontology, and converts them to the local ontology of the data source that is queried. The query results are converted in the opposite direction. Before a data access service handles a query, it checks whether or not the user is authorized to access the data source by contacting the authorization server. The data access services handle SPARQL queries from the semantic mediator. Additionally, they may also be contacted directly by the workflow enactor. This is the case for retrieval of image and assay files, which do not require semantic mediation. The requested data is typically not returned to the workflow enactor, but delivered to file at a specified temporary storage location. The workflow enactor receives the unique identifiers for files that have been created, which it can forward to the data-mining service so that the latter can retrieve and analyse the data.

Fig. 1. The data analysis architecture of ACGT
There are two relevant aspects with regards to the terminology we use. First of all, we use the term “data access service” to refer to a class of services, e.g. the DICOM data access service, as well as for referring to specific instances, e.g. the data access service for DICOM database X. The distinction should always be apparent from the context. Secondly, each data access service is not actually a standalone web service. Within the OGSA-DAI framework multiple data access services are deployed as different data resources within a single OGSA-DAI web service. This has implications for the addressing of the data access services, but is not important for the remainder of this paper.

**Query functionality**

For the implementation of the query functionality for relational databases it is necessary to translate queries from SPARQL to SQL. For this, we are using the Open Source package D2RQ [17]. It can wrap a relational database into a virtual, read-only Jena RDF graph [18], rewrite SPARQL queries and Jena API calls into application-datamodel-specific SQL queries, and transform the returned data into RDF triples. We therefore only had to integrate this functionality into the OGSA-DAI activity framework.

Realizing a data access service for medical image databases requires more effort. First of all, custom code is needed to implement the query translation. As the DICOM information model maps naturally to RDF, it is relatively straightforward to express DICOM queries in SPARQL. However, the DICOM standard only provides limited query functionality, which means that only a subset of syntactically valid SPARQL queries can be expressed as DICOM queries. For the initial implementation, we only support SPARQL queries that can either be directly converted to a DICOM query, or that can be handled using a single DICOM query combined with filters at the data access service that do not require temporary storage of query results (i.e. any query match that is returned by a DICOM server is either immediately discarded, or after optional conversion, immediately returned to the client). This way, the data access service does not need to store intermediate results, and implementation is significantly simplified. Figure 2 shows an example of a supported SPARQL query for a DICOM image repository.

For the medical image data access service, image retrieval functionality was also added; the ability to query the image metadata is of limited use if the actual images cannot be retrieved. The retrieval functionality has been implemented using OGSA-DAI’s activity framework so that it can be invoked in various ways. For example, a single request message can be used to query the image metadata, and to asynchronously retrieve and deliver the corresponding images.
Our third data access service provides access to the BASE database, a database for storing the results of microarray analysis [19]. The data access service interacts with the BASE database by way of a Web Service interface. The current implementation of the data access service provides retrieval of assay files, given their unique identifiers. More advanced query functionality is not provided, as this has not been needed yet. Typically assay files are obtained by first querying the clinical data, e.g. for all patients with an ER-negative tumor that responded positively to treatment, and next retrieving the corresponding assay files from BASE.

**Miscellaneous functionality**

Due to the heterogeneity of the data sources, each data access service requires code that is specific to its type of data source. However, the different data access services also need to provide common functionality, which offers the opportunity for code reuse. The main mechanism by which the OGSA-DAI platform encourages the reuse of code is through its activity framework [14]. Requests from clients to an OGSA-DAI data resource can contain multiple activities, linked together into a pipeline. For example, the first activity may comprise a query to a DICOM server for a set of image identifiers. A second activity may extract the identifiers and retrieve the corresponding images. The images may be fed to a third activity, which compresses the image data, and feeds the resulting archive file to a fourth activity, which delivers the archive to a specified FTP server. Al-
though the interface of the query activities for the different data access services is typically the same (thus providing a homogeneous interface), their implementation is typically highly dependent on the type of data source that is queried. Activities further in the pipeline are typically more generic and their implementation may be reused by multiple data access services. The OGSA-DAI platform comes with a large set of generic activities, but we developed additional ones for use by our data access services. One example is an activity for delivering files to the Gridge Data Management System [20], which we use for temporary storage, using myProxy certificates for authentication. Another activity can calculate checksums for data streams. It can be used for testing service functionality after changes to the implementation, as well as for carrying out periodic liveness tests of a running service. We also extended the default ZIP activity so that it can pack multiple files into a single archive.

Integration experiences

The realization of the data access services requires integration of a large number of third-party software libraries. There are two reasons why a large number of third-party packages is needed: firstly, the complexity of the software stack associated with (grid-based) Web Services, the interface standard chosen in ACGT, and secondly, the heterogeneity of the underlying databases, which typically each have their own sets of standards and APIs associated with them. The software stack for the data access services consists of the following layers:

- data access services
- OGSA-DAI
- Globus
- Tomcat

The lowest layer is the Tomcat web service container, which hosts the web services. Globus sits on top of Tomcat; it is used for implementing the certificate-based security framework. The layer above that consists of OGSA-DAI. It provides a modular, activity-based data access framework for use by the layer above it. The top layer consists of the data access services which handle query and result transformation, and data retrieval and storage for the supported data sources. Each class of data access service depends on various third-party libraries for its implementation. For example, the relational data access service uses D2RQ [17] to translate SPARQL queries to SQL, which in turn uses Jena [18]. The DICOM data access service uses Jena as well, together with dcm4che [21] for accessing DICOM servers. The BASE data access service uses client-code provided by the BASE developers for accessing their BASE web service.

Given this setup, one of the biggest problems is managing the dependencies between all different third-party software packages that are used. This is especially
challenging because all data service resources are deployed within the same
OGSA-DAI instantiation and third-party packages (deployed as Java jar files) that
are needed by only one or a few of the data access services are visible to all. This
can lead to dependency conflicts between data access services that are otherwise
independent. A pair of data access services that can individually be deployed suc-
cessfully inside an OGSA-DAI instantiation may not necessarily be successfully
deployed alongside each other.

Three concrete issues that we encountered may help to illustrate the types of in-
tegration problems this gives. Firstly, after we had discovered and reported a bug
in a third party library we used (Jena), we could not deploy the release that in-
cluded the fix, as this new release was incompatible with another third party li-
brary (D2RQ) that we were using.

Secondly, we have experienced problems deploying compiled and packaged
code provided by other partners in ACGT, which was due to a slight incompatibil-
ity in an underlying third-party library (Axis) provided by the version of the
(Globus-based) web service container that was used. Fortunately, the incompati-

dility did not exist at the source code level, so rebuilding the code with the third-
party libraries of the container where the service was to be deployed fixed the
problem.

We encountered a third problem after we had upgraded the web services con-
tainer, which was required to fix a dependency conflict. One of our services would
now hang when handling requests. As it turned out, this was due to a change of the
third-party library implementing the JavaMail API, which resided three layers be-
low our code. It was due to a more strict implementation of the JavaMail API,
which in turn revealed a bug in another third-party library (Axiom), which relied
on a more lenient interpretation of the API’s contract in order to function cor-
rectly.

It is worth pointing out that in all three cases, the fact that source code was
available for all third-party software components greatly helped in tracking down
and solving the problem.

Discussion

We have implemented OGSA-DAI data access services for three types of data
sources: relational databases, medical image databases and a micro-array database.
The main research question is how to best provide a syntactically homogeneous
interface, and a key question is the query language that is used. We have chosen
SPARQL as the common query language and have demonstrated that it can be
successfully applied to relational databases and DICOM image databases.

For the relational databases, the SPARQL language does not support all fea-
tures offered by the query language of the data source, SQL. For instance, it does
not support aggregation of data (averaging, summation, counting, etc). So aggre-
gation needs to be performed at the client-side, even though the underlying database supports it directly, which negatively affects performance. The actual use of the system by the end users will clarify whether this is a problem that needs to be addressed.

For medical image databases, SPARQL is more expressive than the query support provided by the DICOM protocol. For this reason, the data access service does not support all queries. These limitations are currently described as text, but should be expressed in a more formal manner, so that other services and applications can interpret these and handle accordingly. In order to select a suitable formal framework for this, we need to thoroughly review the capabilities and limitations of all relevant data sources.

A capability-restricted data access architecture has the advantage that it is easier to develop data access services for data sources; as a consequence, new data sources can be integrated much more quickly. It may, however, complicate applications and services that use the data access services. A higher level data access service may therefore be introduced that hides query restrictions of the underlying services. This generic service would decompose queries for a specific data access service as need be, store the intermediate results, and join these to produce the final answer. This would facilitate implementation of the semantic mediator, while incurring a slight performance penalty. However, this higher-level data access service may also carry out generic optimizations such as caching of query results, resulting in performance gains.

Another open issue is how to provide text-based query functionality. There are many public biomedical databases where part of the data is free text. Examples are descriptions of microarray experiments (e.g. in GEO [22] and ArrayExpress [23]), descriptions of gene and protein functions (e.g. in UniProt [24] and EntrezGene [25]), and abstracts and titles of medical publications (e.g. in PubMed [26]). Although most databases provide keyword-based functionality for querying data, this method of searching is not directly supported by SPARQL, so it is not immediately obvious how to extend the current data access services interface to support this functionality. One approach would be to add a separate text-based query interface for data sources that support this. This exposes more details of the underlying data source, resulting in a less homogeneous interface. This is undesirable but may be unavoidable in practice. However, there is a more important question that needs to be answered first: how should querying of text data be handled by the semantic layer? This is an important question as it determines the query interface that is available to end-users, but answering it falls outside the scope of this paper.

To give an impression of the overhead caused by the use of data access services, compared to direct interaction with the databases, we can report the results of performance experiments that we have carried out. The amount of overhead depends on various factors, including the complexity of the query, the amount of results that are returned, and the underlying database. For simple queries the performance degradation could be as much as a factor hundred (in particular for the re-
lational database, which responds very quickly). For more complex queries the overhead decreased significantly, down to a factor of two (for the DICOM database). Overhead was similarly low for retrieval of bulk image and microarray data, but high for retrieval of bulk data that is returned in the XML response message. The latter is due to limitations of the API for constructing the response message, which needs to be constructed entirely in memory before it can be sent to the client.

Finally, many of the problems encountered when deploying data access services for heterogeneous data sources are of a practical nature. For reasons of scalability, all data access services are deployed in the same web services container. This implies however that they run inside the same virtual machine, which can lead to unexpected conflicts. The complexity of the grid-based web services stack in combination with the need to use many third-party libraries, each with their own dependencies and particular implementations of part of the web services stack, makes it a challenge to resolve dependency conflicts.

Acknowledgments  This research has been supported by the Advanced Clinico-Genomic Trials in Cancer (ACGT) project (FP6 IST-2005-026996) funded by the European Commission. Thanks to the anonymous reviewer for the useful feedback.

References
Unification of heterogeneous data towards the prediction of oral cancer reoccurrence

Konstantinos P. Exarchos$^{1,2}$, Yorgos Goletsis$^3$, Dimitrios I. Fotiadis$^1$.

$^1$ Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, Ioannina, GREECE
$^2$ Dept. of Medical Physics, Medical School, University of Ioannina, Ioannina, GREECE
$^3$ Dept. of Economics, University of Ioannina, Ioannina, GREECE

Abstract. Oral cancer is the predominant neoplasm of the head and neck. Annually, more than 500,000 new cases of oral cancer are reported, worldwide. After the initial treatment of cancer and its complete disappearance, a state called remission, reoccurrence rates still remain quite high and the early identification of such relapses is a matter of great importance. Up to now, several approaches have been proposed for this purpose yielding however, unsatisfactory results. This is mainly attributed to the fragmented nature of these studies which took into account only a limited subset of the factors involved in the development and reoccurrence of oral cancer. In this work we propose a unified and orchestrated approach based on Dynamic Bayesian Networks (DBNs) for the prediction of oral cancer reoccurrence after the disease has reached remission. Several heterogeneous data sources featuring clinical, imaging and genomic information are assembled and analyzed over time, in order to procure new and informative biomarkers which correlate with the progression of the disease and identify early potential relapses (local or metastatic) of the disease.

Keywords: oral cancer, dynamic Bayesian networks, reoccurrence, disease modeling

Introduction

Oral cancer refers to the cancer that arises in the head and neck region, i.e. in any part of the oral cavity or oropharynx. Oral cancer constitutes the eighth most
common cancer in the worldwide cancer incidence ranking; more than half million patients are diagnosed with oral squamous-cell carcinoma worldwide every year [1]. Oral cancer is highly related to the sex of the patient, with men facing twice the risk of being diagnosed with oral cancer than women. Research has revealed several risk factors associated with the development of oral cancer. Smoking and excessive consumption of alcohol, and especially the combination of the two, constitute predominant risk factors for developing oral cancer. Moreover, sun exposure is another important risk factor, particularly for the cancer of the lip [1]. Some studies have also suggested that infection with the human papillomavirus (HPV) is associated with oral cancer, especially with occurrences in the back of the mouth (oropharynx, base of tongue, tonsillar pillars and crypt, as well as the tonsils themselves) [2].

Cancer cells can spread to other adjacent parts of the neck, the lungs or elsewhere in the body. A common metastasis occurs in the neck lymph nodes through the lymphatic system which helps the cancer cells spread. Although nowadays, the continuous improvements in treatment protocols of cancer have achieved high rates of successful disease disappearance [3], there is a critical stage for the disease evolvement after the treatment called remission; during this stage there is no clinical, laboratory or imaging evidence of the neoplastic mass and the patient is considered cancer free. Nevertheless, even at this point some “invisible” disease particles might still be present leading to a potential spread or metastasis of the disease. Specifically, in terms of oral cancer, locoregional recurrence rates after the disease has reached remission have been reported in the range of 25-48%; such high figures can be justified given the deeply infiltrative nature of these tumors, as well as, the significant potential for occult neck metastasis [4].

The recurrence rates for oral cancer are quite high and they also suffer from poor prognosis, which can be partly attributed to histologically unfavorable features [4]. Moreover, patients with oral cavity cancer have to deal with the impact of the disease and its treatment on their physical appearance and on the ability to eat and speak, and subsequently with a significant decrease of the quality of life. Hence, early detection of reoccurrence might prove very beneficial [5]. Currently implemented methods aiming to predict oral cancer reoccurrence after the disease has reached remission, have reported quite inadequate results. Although several factors have been associated with the reoccurrence of oral cancer, such as age, site and stage of the primary tumor as well as histological features, they have not been studied altogether in a collective study. Moreover, especially in the molecular basis of the disease, currently available biomarkers are limited in number and efficiency [6, 7]. The efficient combination of the already known ones will greatly benefit the accurate stratification of the patients in terms of staging.

In the general framework of disease prognosis and modeling, several diverse approaches have been proposed in the literature. Most of them involve a prognostic model which implements a risk score depicting the progression of the disease and the general condition of the patient. Based on this score, simple decision rules are used to stratify the patients into several risk categories [8, 9]. More recent ap-
approaches utilize advanced machine learning algorithms, such as Artificial Neural Networks (ANNs) or Support Vector Machines (SVMs) which accept as input several variables and provide prediction about the desired outcome. However, most of these approaches use a “black-box” architecture and thus do not provide adequate reasoning about the decision [10, 11]. In addition, it is very cumbersome, if not infeasible to represent properly temporal problems using these algorithms. These issues pose significant limitations for the acceptability of the produced decision systems both by the medical community and the patients. In the case of oral cancer, and cancer in general, the physicians are extremely interested in knowing if, when and why a reoccurrence will appear. Hence, especially for the problem under consideration (i.e. oral cancer reoccurrence prediction) it is very important to provide sufficient justification about the prediction, but also to introduce the time dimension in the modeling procedure.

In this work, we present an efficient framework in order to systematically study and analyze the factors associated with the reoccurrence of oral cancer, after the remission of the disease. This objective involves the integration of heterogeneous clinical, imaging and genomic data, thus facilitating the multiscale and multilevel modeling of the disease progression over time. Due to the constantly evolving nature of the disease, we employ DBNs, which efficiently cope with temporal causalities, thus, identifying the timing of a potential reoccurrence. Moreover, the intuitive design of DBNs allows for comprehensible decisions coupled with adequate justification. The multitude of gathered data is likely to uncover the evolution and development of the disease during remission, thus assisting the monitoring of patients after treatment, but also contribute towards the accurate stratification of patients in terms of staging. Knowing in advance the progression of the disease, i.e. identifying groups of patients with higher/lower risk of reoccurrence is a key factor towards the determination of the most proper treatment.

Materials and Methods

Clinical scenario

In order to clarify the steps of our study, a clinical scenario is employed which is shown in Figure 1. Initially a patient is diagnosed with cancer through traditional clinical procedures. At this point the physician gathers the required data in order to extract the baseline profile and the patient is treated properly. After the physician’s therapeutic intervention, the patient either reaches complete remission or particles of the cancer tissue still remain intact. In the latter case the patients do not qualify for the purposes of our study, whereas from the patients in complete
remission, where the cancer is no longer visible, data are further collected, forming the post-treatment profile. Afterwards, and during a two year time span, data are collected from the patient regularly (i.e. scheduled visits are planned for months 1, 3, 6, 9, 12, 15 and 18 after treatment) in order to formulate as a personalized follow-up signature, which is being constantly analyzed. The choice of the follow-up period was determined by the fact that a reoccurrence is most likely to appear in a two year period after the initial treatment. The purpose of this analysis is to stratify the patients in two clusters: i) low risk of disease reoccurrence and ii) high risk of reoccurrence. Hence, we are able to duly identify relapses of the disease and adjust the follow-up treatment accordingly.

Data collection

The progress of the disease in a total of 150 patients with oral squamous cell carcinoma is evaluated during the present study. The cases are collected from two major clinical centers which reside in Italy and Spain. According to available literature 70-80% of these patients are expected to achieve complete remission of the disease after treatment, and an approximate 30-40% of them will develop a reoccurrence of the cancer. Relapses during a two-year time span are marked, as well as the timing of the relapse, and the patients are grouped in two categories, the relapsers and the non-relapsers, which we aim to discriminate by studying and analyzing a multitude of heterogeneous data.

Figure 1: Clinical scenario employed in our study.
Due to the complex nature of cancer, a major challenge towards its diagnosis and treatment is to formulate a collective approach in order to “frame” every possible aspect. For this purpose we propose a holistic approach which involves the integration and analysis of multiscale and multilevel data. Specifically, clinical, imaging and genomic data are assembled ranging in the scale of dimension and localization. The employment and careful analysis of the above heterogeneous data is likely to reveal the interactions which take place during oral cancer onset and progression. Consequently, the data collected from every patient will comprise the following information:

- Clinical data from health records and standard laboratory markers, histological data from tumor mass specimen
- High throughput genomic data from tumor tissue specimens and circulating cells, profiling gene expression at whole genome level by oligo-RNA microarrays
- Imaging data of the prime tumor mass (and secondary localizations if present)

All these data will be efficiently integrated into a single repository formulating the basis of our study. The data involved in the present study along with the specific techniques employed for their manipulation and analysis are described in detail in the sections that follow.

**Clinical**

For the diagnosis and monitoring of patients with oral cancer the following types of clinical data are assembled:

- Anamnesis
- Demographics
- Risk factor
- Tumor clinical aspect
- TNM staging
- N characteristics

Anamnesis refers to the detailed medical review of the patient’s past health state. Detailed information about the patient’s past health problems, general health state, family medical history, oral cancer risk factors and symptoms is gathered in order to establish the diagnosis. Demographic data along with several risk factors are also assembled in order to aid the diagnosis. Next the tumor’s clinicopathological stage and developmental phase are evaluated. The most common staging system used for oral cancer is the TNM system. Moreover, several markers have been proven to affect the patient’s response to adjuvant and neo-adjuvant treatments [12, 13]. In the present study we compile an extensive list containing all these clinical factors in order to perform a collective study of their relation with
oral cancer progression and treatment efficacy. All these data, which comprise the clinical data associated with oral cancer, are thoroughly analyzed for the purposes of the present study.

**Genomic**

Current advances in the field of genomics have enormously facilitated the thorough analysis of gene expression within cells and tissues. Hence, we are able to extract important information about the interactions and biological pathways which take place during cancer evolution. The framework of the present work employs oligonucleotide and complementary DNA arrays in order to unravel the molecular basis of oral cancer. Nucleic acid arrays have rapidly become a popular investigational tool for cancer biologists, towards the identification of robust genetic biomarkers, thus, shedding considerable light into the complexity of the disease. Systematic analysis of gene expression data is likely to yield potential tumor markers, or reliable combinations of biomarkers, that can be afterwards used in the daily practice for the diagnosis and monitoring of carcinoma of the head and neck.

Gene expression data come from a feature extraction (FE) file. An FE file is a tab delimited text file comprising of expression values (Log2-ratio data), raw intensity data, background information, metadata regarding the experiment and the scanning settings, gene annotation, etc. A typical FE file is shown in Figure 2.

![Figure 2: Typical entities extracted from a microarray experiment.](image)

In the present study, all microarray experiments are conducted using the same platform, the same array design and the same FE procedure, in order to minimize the risk of possible sources of variability in the data, other than biological variability.
Especially for genomic data, a preprocessing stage is necessary for enhancing the quality of the data. After obtaining the gene expression data from the microarray experiments the duplicate and control features are eliminated. Control features are negative and positive control elements usually represented by empty features or spots that are hybridized independently from the original sample. Whereas, duplicate features are probes corresponding to a gene or a known internal control sequence which are printed more than once in the array, usually in random positions. They are used to verify the internal consistency of the data and the regional quality of the hybridization. Furthermore, data with high variability, too low signal and genes with a large number of missing values, constituting unreliable expression levels are carefully filtered out.

The overall flowchart for the basic preprocessing of the gene expression data is shown in Figure 3.

![Flowchart](image)

**Figure 3:** Preprocessing of the gene expression data.

**Imaging**

Image data from the cancerous tissue can reveal certain significant characteristics of the localization and progress of the disease. The present study employs
MRI and CT images. The manipulation of the employed images involves the following main steps, which are also depicted in the flowchart of Figure 4.

- Image preprocessing
- Definition of regions of interest (ROIs)
- Extraction and selection of features
- Classification of the selected ROI

Initially, the images need to be preprocessed properly in order to improve their quality to facilitate the overall image analysis procedure. The most common types of imaging data contamination are noise and artefacts. Noise causes random distortion in the data, and although several approaches have been proposed in the literature (e.g. application of filters), it is still quite difficult to remove it, due to its random nature. On the other hand, artefacts usually involve more deterministic perturbations of the data, hence it is easier to detect and omit them. These problems can be attributed to several factors such as human error, measuring device limitations, etc. Other types of image preprocessing involve edge enhancement (e.g. unsharpening, wavelet transformation), image contrast enhancement (histogram equalization) and image standardization.

In the next step, we detect regions of interest (ROIs), i.e. regions of the preprocessed image bearing enhanced role for our purposes. For the initial approximate definition of some ROIs, a specialized radiologist pinpoints sites of interest,

Figure 4: Image data analysis and manipulation.
i.e. tumor center, lymph nodes or potential infiltrations. Moreover, automatic methods are also be employed for the detection of ROIs. Active contour models are often employed for automatic definition and tracking of anatomical contours in 2D medical images due to their ability to approximate accurately the random shape of organ boundaries. Seeded region growing is another example of semiautomatic method widely used for the definition of ROIs in medical images.

Afterwards, several features are extracted from the ROIs in order to uniquely characterize the image itself or structures contained in the data. Some of these features represent quantitative measurements with certain physical meaning, that a specialized physician must take into account in order to formulate the diagnosis. However, in some cases features with no apparent physical meaning can be extracted due to their enhanced discriminative potential. The most common features employed for the analysis of medical images are: pixel based features, texture features, shape features (transformation dependent and transformation independent). Specifically, in the present study, the following features are calculated from each ROI:

- Six (6) features from first order statistics
- Forty eight (48) features from spatial gray-level dependencies matrix
- Twenty (20) features from gray-level differences matrix
- Twelve (12) features from Law’s texture energy measurements and
- Three (3) features from fractal dimension measurements

Additional features describing specific properties of the image under consideration are assessed, such as tumor volume, periosteal infiltration, etc. All features extracted during this stage are deposited in a collective repository along with the genomic and clinical data.

**Dynamic Bayesian Networks (DBNs)**

In the present study we employ DBNs in order to early identify potential relapses of the disease, during the period of remission. As it is described in the clinical scenario, a snapshot of the patient’s medical condition is acquired during every predefined follow-up with the doctor. By exploiting the information of history snapshots we aim to model the progression of the disease in the future. The proposed prognostic model is based on DBNs, which are temporal extensions of Bayesian Networks (BNs.) [14]. A BN can be described as $B = (G, P)$ where $G$ is a directed acyclic graph, where the nodes correspond to a set of random variables $X = \{x_1, x_2, ... , x_N\}$, and $P$ is a joint probability distribution of variables in $X$, which factorizes as:

$$P(X) = \prod_{i=1}^{N} P(x_i \mid \pi_G(x_i))$$  \hspace{1cm} (1)
where $\pi_G(x)$ denotes the parents of $x$ in $G$. A DBN can be defined as a pair $DB = (B_0, B_{\text{trans}})$ where $B_0$ is a BN, defining the prior $P(X_0)$ and $B_{\text{trans}}$ is a two-slice temporal BN (2TBN) which defines $P(X_t | X_{t-1})$. The semantics of a DBN can be defined by “unrolling” the 2TBN until we have $T$ time-slices. The resulting joint distribution is given by:

$$
P(X_1, X_2, ..., X_T) = \prod_{t=1}^{T} \prod_{i=1}^{N} P(x'_t | \pi(x'_t))$$

In order to build a model that successfully evaluates the current state or predicts a state in the future (next time slice), we need to train both the structure of the DBN ($G_0, G_T$) and the parameters of the conditional probability distributions, using both expert knowledge as a prior model and experimental data to get a more accurate posterior model. After the training procedure, we obtain a model as the one shown in Figure 5. By providing some evidence to the model, we are able to compute the probability of any variable for every time slice (i.e. in any predefined follow-up visit), including of course the probability for reoccurrence.

![Figure 5: Provisional architecture of the employed DBN model.](image)

For the development of the DBN two implementations have been explored. In the first implementation every source of data is used separately, in order to build a distinct DBN, specifically tailored for a certain type of data. Consequently, three DBNs are developed and their outputs are combined using a meta-classification function (Figure 6(a)). In the second, all sources of data are employed altogether...
in order to develop a single DBN (Figure 6(b)). However, in both implementations, the contribution and feedback from a specialized doctor, during the DBN construction, is substantial. The two implementations are depicted in Figure 6.

![Diagram](image)

**Figure 6:** Analysis schemes: (a) multiple DBNs, (b) single DBN.

As this work is currently under development, detailed testing of both implementations depicted in the figure above is needed so as to assess the potential of each one. The assessment will be done using an annotated dataset, covering the two-years follow-up data, which is currently being populated.

**Discussion & conclusions**

In the present study we propose an advanced framework which implements heterogeneous sources of data towards the prediction of oral cancer reoccurrence in patients that have reached remission. A large amount of clinical, genomic and imaging features are analyzed in order to extract biomarkers that are highly associated with relapses of oral cancer. Thus, we overcome a major limitation of similar studies in the field that employ only a confined subset of features that are associated with oral cancer. Another significant challenge is to capture the disease progression over time. For this purpose we employ DBNs, which are specifically designed to represent temporal causalities. The inclusion of the time dimension is very important as most doctors are interested – even with a rough approximation – in the timing of the reoccurrence. Furthermore, DBNs are able to provide reasoning for the reported decisions, thanks to their transparent architecture. This characteristic is very appealing, if not prerequisite by the medical community. Hence, not only we are able to predict a certain outcome but also to gain insight about the rationale of every decision. In overall, the currently proposed framework contrib-
utes significantly towards the monitoring of oral cancer evolution since it can answer if, when and why a reoccurrence might appear.

Acknowledgements

This work is part funded by the European Commission NeoMark project (FP7-ICT-2007-224483) – ICT enabled prediction of cancer reoccurrence.

References

Building a System for Advancing Clinico-Genomic Trials on Cancer

Stelios Sfakianakis, Norbert Graf, Alexander Hoppe, Stefan Rüping, Dennis Wegener, Lefteris Koumakis, and George Zacharioudakis

Abstract The analysis of clinico-genomic data poses complex computational problems. In the project ACGT, a grid-based software system to support clinicians and bio-statisticians in their daily work is being developed. Starting with a detailed user requirements analysis, and with the continuous integration of usability analysis in the development process, the project strives to develop an architecture that will substantially improve the way clinico-genomic trials are conducted today. In this paper, results of the initial requirements analysis and approaches to address these requirements are presented. We also discuss the importance of appropriate metadata to tailor the system to the needs of the users.

1 Introduction

The goal of the Advancing Clinico-Genomics Trials on Cancer (ACGT) project is to develop an open-source and open access IT infrastructure that provides the biomedical research community with the tools needed to integrate complex clinical information and make a concrete step towards the tailoring of treatment to the patient[3]. The necessity of such an environment is evident today more than ever due to the recent advancements in

Stelios Sfakianakis · Lefteris Koumakis · George Zacharioudakis
Institute of Computer Science, FORTH, Greece, e-mail: ssfak@ics.forth.gr

Norbert Graf · Alexander Hoppe
University Hospital of Saarland, Paediatric Haematology and Oncology, D-66421 Homburg, Germany, e-mail: {norbert.graf,alexander.hoppe}@uniklinikum-saarland.de

Stefan Rüping · Dennis Wegener
Fraunhofer IAIS, Schloss Birlinghoven, 53754 St. Augustin, Germany, e-mail: {stefan.rueping,dennis.wegener}@iais.fraunhofer.de

1 http://www.eu-acgt.org
high throughput genomics and post-genomics technologies. These technologies yield an enormous pool of data that needs to be managed, analysed, correlated, and comprehended for the treatment of diseases like cancer and for the benefit of the community at large.

In this paper we discuss how the clinical requirements for such an environment can be addressed in a large-scale system and how appropriate meta data can be used to achieve satisfaction of the end user. The rest of the paper is structured as follows: Section 2 gives an overview over the user requirements analysis that was conducted in ACGT and highlight the main challenges. Section 3 introduces the main ACGT architecture, and in Section 4 some new approaches to address the user requirements are presented. Section 5 concludes.

2 The End-User View

Treatment and survival of patients with cancer is increasing steadily for most age groups as shown in Fig. 1 which gives the average annual percentage change over a period of 10 years. One of the most important reasons for this success story is the enrollment of patients in prospective clinical trials. Nevertheless, for most clinical trials in cancer, the number of patients recruited is much lesser than the number of eligible patients. In adults only 5% of cancer patients are participating in such trials. Therefore, higher rates are of utmost importance, especially in those cancers with still a dismal prognosis. To achieve this goal it is necessary to facilitate the building and running of clinical trials and to attract more patients to participate. In addition, the improvement in molecular biology has to be taken into account to create more clinico-genomic trials.

Recent advances in methods and technologies in molecular biology have resulted in an explosion of information and knowledge about cancer and its treatment. As a result, our ability to characterize and understand the various forms of cancer is growing exponentially. Information arising from post-genomics research and combined genetic and clinical trials on one hand, and advances from high-performance computing and informatics on the other, is rapidly providing the medical and scientific community with an enormous opportunity to improve prognosis of patients with cancer by individualizing treatment. To achieve this goal, a unifying platform is needed that has the capacity to process this huge amount of multi-level and heterogeneous data in a standardized way. Multi-level data collection within clinico-genomic trials and interdisciplinary analysis by clinicians, molecular biologists and others involved in life science is mandatory to further improve the outcome of cancer patients. It is essential to merge the research results of biomolecular findings, imaging studies and clinical data of patients and to enable users to easily join, analyze and share even great amounts of data. To provide a functional
and user-friendly platform it is of utmost importance that the development of such a platform is user-driven and evaluated by end users right from the planning and development phase. Tools and software developed within ACGT are based on the user’s needs and have to be in accordance with ethical and legal requirements of the European Community.

The project has selected indicative Clinical Trials on Cancer, namely breast cancer, pediatric nephroblastoma and in-silico modeling and simulation of tumor growth and response to treatment, for the initial requirements gathering activity. Since ACGT sees the requirements engineering process as a structured set of activities which will lead to the production of the final system requirements, an iterative requirements engineering process has been adopted, mainly based on scenarios and prototyping. Inputs to the requirements engineering process are information about existing systems, user and stakeholder needs, organizational standards, regulations and other domain information. As clinico-genomic trials are in the center of ACGT a Clinical Trial Management System is of utmost importance, to collect clinical, biomedical, imaging and other trial specific and relevant data. ACGT will provide such a tool, called ObTiMA [5, 6], whose functionality includes administrative and scientific aspects of clinico-genomic trials.

It has to be stressed that such a complex platform as ACGT, dealing with extremely sensitive data (patient data) and used by many different, sometimes multi-role, end-users, having different needs and requirements, a Data Protection Framework for ACGT is mandatory. This is based on the anonymization of patient data, the informed consent from participating patients and the binding of partners/centers by contracts to the ACGT policies and procedures and will ensure compliance with the Data Protection regulations. In addition and from an ethical point of view it is strongly demanded to

Fig. 1 Average annual change in survival in patients with cancer [4].
let patients participate in and have a measure of influence over the processing of their genetic data.

To assure the success and functionality of the ACGT environment end-users are involved in every step of the development process.

### 2.1 Main Requirements

In summary, the user requirements that have been identified by the clinical experts for the ACGT environment can be divided into the following aspects:

- **Appropriateness**: the data analysis environment should provide the appropriate tools and services to support users in the state-of-the-art scientific analysis of biomedical data. Section 4.1 introduces the use of the GridR component [7], a “gridified” version of the well-known R statistical software, which is a de-facto standard for many kinds of biomedical data analysis.

- **Extensibility and reusability**: the platform should be easily extensible to new tasks and existing solutions should be easily reusable and transferable to similar problems. Extensibility is addressed in Section 4.1, while an important aspect of reusability, namely quality control, is described in Section 4.5.

- **Performance**: the system must be performant enough to facilitate large analysis and optimization tasks, which calls for an efficient use of the grid architecture. Challenges exists not only because of the size of the data sets (see Section 4.2), but also from their complexity and heterogeneity (Section 4.3), which is a result of the distributed nature of pan-european clinical trials.

- **Security**: The system must be secure and protect the privacy of the involved patients. This is discussed in Section 4.4.

- **Usability**: the system should be easy to use for inexperienced users, but also provide a powerful interface for experts. Usability is best achieved by a continuous process of evaluation and optimization. In Section 4.6, approaches to automatically identify parts of the system that require a high amount of attention are discussed.

### 3 The ACGT Architecture

The complexity and the diversity of user requirements have a strong impact on the design of the ACGT architecture. It is evident that a multidisciplinary and multiparadigm approach is necessary in order to deal with these requirements. For these reasons the ACGT platform is designed according to the
following technologies and standards: Service Oriented Architecture (Web Services), the grid, and the Semantic Web. In essence, the grid provides the computational and data storage infrastructure, the general security framework, the virtual organization abstraction and relevant user management mechanisms etc. The machine to machine communication is performed via XML programmatic interfaces over web transport protocols, which are commonly referred as Web Services interfaces. Finally the Semantic Web adds the knowledge representation mechanisms through the means of OWL ontologies, the implementation-neutral query facilities with the SPARQL “universal” query language and the associated query interfaces.

The adopted architecture for ACGT is shown in Fig. 2. A layered approach has been followed for providing different levels of abstraction and a classification of functionality into groups of homologous software entities. In this approach we consider the security services and components to be pervasive throughout ACGT so as to provide both for the user management, access rights management and enforcement, and trust bindings that are facilitated by the grid and domain specific security requirements like pseudonymization. Apart from the security requirements, the grid infrastructure and other services are located in the first (lowest) two layers: the Common Grid Layer and the Advanced Grid Middleware Layer. The upper layer is where the user access services, such as the portal and the visualization tools, reside. Finally, the Bioinformatics and Knowledge Discovery Services are the “workhorse” of ACGT and the corresponding layer is where the majority of ACGT specific services lie.
4 Addressing the User Requirements in ACGT

In the following, we will try to give a short overview of how to integrate the user requirements of Section 2 into the ACGT grid architecture.

4.1 Extensibility

The requirement for extensibility is very important in the context of grid-enabled data mining [9]. Especially, in order to keep track with new scientific developments, it is crucial to be able to quickly integrate new analysis services or algorithms into a data mining platform. Related to the ACGT environment, extensibility denotes the possibility of extending the environment at the workflow level, at the service level, or at the algorithm level. In order to deal with such requirements we have found that the use of metadata descriptions and the ontology based integration of the ACGT platform components provides a future proof approach to extensibility. In the following we will introduce GridR [7] as an example to demonstrate how the ACGT system can easily be extended by new services and algorithms.

GridR is an analysis tool based on the statistical environment R [2] that allows using the collection of methodologies available as R packages in a grid environment. The aim of GridR is to provide a powerful framework for the analysis of clinico-genomic trials involving large amount of data (e.g. microarray-based clinical trials). The GridR service (see Fig. 3) combines the wide spectrum of methods available in R with an effective distributed grid data management system (DMS) and efficient execution supported by a grid resource management system (GRMS), see [8]. In this fashion, users can make efficient use of distributed, parallel computational resources in their R scripts, while all the technical details are hidden from them. The R code to be executed can be given directly by the user in the form of a script, but in order to increase the possibility of distributing and re-using code, the intended way to execute R code is by storing it in a metadata repository, such that it becomes available to the whole system. Technically, an R function or script f thus becomes an f-service. Consequently, users who prefer to work on the workflow level and not edit their own code can make use of available R scripts and even all the single R functions in R libraries in their workflows.

Along these lines new algorithms can be “gridified” and be seamlessly integrated with the rest of the ACGT grid environment without a need for changing the service’s or the R script’s implementation.
4.2 Large Data Sets

Data is the most valuable asset of ACGT and therefore the platform should be able to manage big data sets in an efficient and secure way. The storage and the transfer of the data is of particular importance and something that should be taken care in a uniform way in the whole ACGT environment.

Data storage requirements are addressed by the grid infrastructure and the ACGT “Data Grid”, which controls the sharing and the management of large amounts of distributed data. However, an additional issue has to do with the protocols, infrastructure, and policies for moving these large data sets to the processing nodes where the data analysis is performed. In some cases the grid infrastructure could be employed so that instead of moving the data around, the processing tasks, by the means of grid job submission and scheduling services, are transferred where the data reside. Nevertheless, the majority of services and data processing tools in ACGT are implemented as XML Web Services that are accessible through the network. Being a text format, XML is well known for its unfriendliness for transferring binary data. There are a couple of solutions for this ranging from encoding the binary data in hexadecimal or, most often, in Base64 text format, to using “attachments” in the SOAP messages. Nevertheless these approaches impose additional processing and bandwidth costs and so we opt for another option, which is to transmit references to data as part of the Web Services interaction while the data itself can be transferred through “out of band” channels, e.g. by the means of GridFTP. This approach offers the advantage of “quicker” XML interactions, easier and more performant service composition since there is no need to “get” (download) a huge binary data set in order to “give” (upload) it to another service, identity of the data so that they can associated with metadata through their references, etc.
4.3 Complex Data

A particular characteristic of data analysis in clinical trials is that the data used in a statistical analysis can be very heterogeneous and dynamic, meaning that many different tools and approaches may be necessary to analyze the data, but also that intermediate results may become invalid as the trial progresses and more data becomes available. The situation is further exacerbated when improved interactivity can result in changing workflows on the fly and cloning running workflows to explore alternatives in parallel. This involves the risk that the user becomes overwhelmed by the enormous amount of information and choices that are available to him. Hence, approaches to help the user better deal with the possibilities of the system are necessary.

For these reasons, a hyperlinked presentation of information has been proposed as a tool for better supporting the collaboration in scientific communities [10]. In essence, provenance information can be viewed as a graph of services invocations, with edges representing several types of lineage and provenance. For example, relationships such as “produced-by”, “part-of”, “derived-from”, “input-of” etc. can be modeled this way. Each entity (e.g. service, data) is identified by an HTTP URI to provide identification, retrieval, and linking facilities for constructing a web of data and metadata in accordance with the Semantic Web vision [11]. Therefore in ACGT we aim to employ the semantic web technology in order to facilitate the tasks of both the users and the intelligent knowledge extraction services. Users are able to navigate to the information graph formed by the casual and other relationships between and among services and data just by following the hyperlinking paradigm that was popularized by the World Wide Web. On the other hand, semantic web enabled software entities are empowered to take advantage of the semantically rich content and to draw conclusions and knowledge based on the referenced ontologies.

4.4 Security

The sensitivity of the patient data requires a strong security framework to provide enough safety nets in order to maintain privacy, confidentiality, and integrity. The grid middleware already supports much of the necessary infrastructure, in terms of certificate based Grid Security Infrastructure (GSI), the Virtual Organization (VO) abstraction and the user credential management, and the Grid Authorization Services (GAS). In ACGT this “system level” security is complemented by “domain specific” mechanisms like pseudonymization that permits the identification of patient specific information without revealing the true person identity. All data is anonymized before their entry in ACGT and even during their analysis all the processing tasks are audited and authorized based on the end users’ identity[1].
An interesting assertion about security of services can be made when using tool repositories as described in Section 4.1: with a standard web service, which can be deployed anywhere, it is principally impossible to give technical guarantees about which code is executed, as only the interface of the service is given and standardized. In general this is a desired property of web services, however, when considering data security, this means that external (legal) measures have to be taken to prohibit the service owner from disclosing information about the data. With the use of tool repositories, it can be guaranteed by a central instance, that the code in the repository has been reviewed and is secure to use, because the code that is being executed is directly transported to the execution site from the repository. In addition, this shipment of algorithms allows to analyze the data on a secure site, without needing to transport sensitive data.

4.5 Quality Control

In dynamic, distributed, and heterogeneous environments with multiple actors and complex use cases it is important to have a continuous validation of the different functional components. Therefore an ACGT validation and testing infrastructure is required to constantly monitor the ACGT services and report any malfunctions. This infrastructure for the automatic testing and validation of ACGT workflows and services is useful both for the initial decision making process about the acceptance of a new service and for the monitoring the status of the ACGT services as a whole. The status of ACGT services and workflows is checked with respect to the following criteria:

- Liveness, i.e. that it’s “alive” and normally operating
- Correctness, i.e. that it delivers the correct results
- Performance, i.e. that it responds in a timely fashion

A number of tests are developed as scripts for each service according to these criteria. These tests are of course service and workflow specific because different components have different notions of correctness or performance. Nevertheless all of them are given some sample input data and parameters and based on this information they validate the target services according to the services’ interface and functionality. The tests are stored centrally and re-evaluated periodically.

The advantage of this testing scenario is that even complex, user-defined workflows can be tested periodically, such that a single user can be notified if a workflow (i.e. a scientific experiment) of hers fails to meet the expected results. In this way, not only software quality, but also the quality of published, clinically relevant findings can be controlled.
4.6 Usability

Much thought in the ACGT project is given to the usability of the final software, including a formal usability analysis and end user integration throughout the runtime of the project to guarantee that the software will meet the requirements of the end users. Usability analysis is an important, but very time consuming process. In this section, we will present some approaches on how to improve the usability of the system using information present in the system’s meta data.

The idea is that workflow execution statistics can be gathered together with other meta data and put into relation with the user’s content with the system. For example, an analysis of workflows which are often canceled can provide the system’s administrator with valuable information on how help users to select a better workflow. A statistic of the execution time of different services can help a developer to choose which services to optimize. A list of often used services can help new users to select good services.

Hyper-linking between meta data, workflow templates and workflow statistics also allow for a more complex reasoning of the users intent. One example could be as follows: the user executes different workflows on a data set, or variants thereof. From the meta data of the data set the system finds out that all the variants of the data set point to the same basic data set (e.g. a trial) and hence can reason that all the workflow executions belong together. It can then search the database of historic workflow executions to see whether a similar groups of workflow have been executed by another user. If this is the case, it is reasonable to assume that both users try to solve a similar problem, and hence the best workflow of the old user can be suggested to the new user. Of course, privacy aspects have to be considered in this kind of scenario.

5 Conclusions

There are a number of projects that aim at developing grid-based infrastructure for post-genomic cancer clinical trials, the most advanced of which are NCI’s caBIG\(^2\) in the USA and CancerGrid\(^3\) in the UK. The overall approach in those projects is somewhat different from the one in ACGT. In caBIG, the bottom-up, technology-oriented, approach was chosen, in which the focus was put on the integration of a large number of analysis tools but with weak concern on data privacy issues. CancerGrid on the other hand addresses the very needs of the British clinical community. In contrast, the goal of the ACGT project is develop a pan-european system that is driven by current demands

\(^2\) Cancer Biomedical Informatics Grid, https://cabig.nci.nih.gov/

\(^3\) http://www.cancergrid.org/
from clinical practice. With two on-going international clinical trials actually conducted in the framework of the project, the approach is top-down, with clinicians’ and biomedical data analysts’ needs at the heart of all technical decisions, considering data privacy issues as central as data analysis needs.

In this user driven endeavor the technical concerns raised by the multiplicity and heterogeneity of user requirements demand state of the art methodologies and technologies. In the ACGT work plan the employment of ontologies and metadata annotations and the realization of intelligent higher level services are the primary implementation targets. Finally, in the realization of this environment, we aspire that the users are also involved. Guided and facilitated by the infrastructure, they can actively participate by creating and sharing information and knowledge. Only this way the ACGT is enriched and improved to become a really useful scientific tool.

Acknowledgments:

The authors wish to thank the ACGT consortium for their contributions and various ideas on which the ACGT project was developed. The ACGT project is funded by the European Commission (FP6/2004/IST-026996).

References


Discovery of Genotype-to-Phenotype Associations: A Grid-enabled Scientific Workflow Setting

Lefteris Koumakis\textsuperscript{1,2}, Stelios Sfakianakis\textsuperscript{1}, Vassilis Moustakis\textsuperscript{1,2}, and George Potamias\textsuperscript{1}\textsuperscript{*}

\textsuperscript{1}Institute of Computer Science, FORTH; \textsuperscript{2}Dept. of Product. Engineering & Management, TUC

Abstract. The heterogeneity and scale of the data generated by high throughput genotyping association studies calls for seamless access to respective distributed data sources. Toward this end the utilization of state of the art data resource management and integration methodologies such as Grid and Web Services is of paramount importance for the realization of efficient and secure knowledge discovery scenarios. In this paper we present a Grid-enabled Genotype to Phenotype scenario (GG2P) realized by a respective scientific workflow. GG2P supports seamless integration of clinico-genetic heterogeneous data sources, and the discovery of indicative and predictive clinico-genetic models. GG2P integrates distributed (publicly available) genotyping databases (ArrayExpress) and utilizes specific data-mining techniques for feature selection – all wrapped around custom-made Web Services. GG2P was applied on a whole-genome SNP-genotyping experiment (breast cancer vs. normal/control phenotypes). A set of about 100 discriminant SNPs were induced, and classification performance was very high. The biological relevance of the findings is strongly supported by the relevant literature.

1 Introduction

Scientific community experiences an increasing need for efficient data management and analysis tools and there is an unprecedented demand for extraction and processing of knowledge. This is more than evident in the domain of bioinformatics since the beginning of the “genomic revolution”. After the completion of the Human Genome Project and the emergence of high throughput technologies (DNA microarrays, high-density SNP genotyping, mass spectrometry etc) a vast amount of biological data are being produced on a daily basis. This has raised the expectation of extracting valuable knowledge for post-genomic personalized dis-

\textsuperscript{*} Corresponding author, potamias@ics.forth.gr
ease treatment. Therefore new challenges for the data analysis and knowledge discovery processes are introduced.

Knowledge Discovery and Data Mining are the most prominent methods and tools for the state of the art scientific discovery. Requirements for biological data management are very demanding due to size and complexity, quality properties (missing values or noisy data are frequent), and inherent domain heterogeneity. These new requirements have given rise to modern software engineering methodologies and tools, such as Grid (Foster 2003) and Web Services (Curbera et al 2002). These new technologies aim to provide the means for building sound data integration, management and processing frameworks.

This paper presents an integrated scenario to support seamless access and analysis of Single Nucleotide Polymorphisms (SNP) genotype data, as produced by relative SNP genotyping platforms. Effort is cast toward the discovery of reliable and predictive multi-SNP profiles being able to distinguish between different phenotypes. The employed data-mining technique is founded on a novel feature selection algorithm. The whole approach is realized in a Grid-enabled scientific (BPEL-compliant – BPEL stands for Business Process Execution Language) workflow editor and enactment environment, and presents an integrated scenario aiming to support Grid-enabled Genotype-to-Phenotype (GG2P) association studies. In particular, GG2P seamlessly accesses and gets phenotypic and genotypic SNP data; analyzes them; and presents results (e.g. the most discriminant and descriptive SNPs) in an appropriately devised html file with links to the Ensembl genome browser.

2 Enabling Technology

With the completion of the human genome and the entrance into the post-genomic era the large amount of data produced makes difficult to extract and evaluate the hidden information without the aid of advanced data analysis techniques. Data mining has successfully provided solutions for finding information from data in many fields including bioinformatics. Many problems in science and industry have been addressed by data mining methods and algorithms such as clustering, classification, association rules and feature selection. In particular, feature selection is a common technique for gene/SNP feature reduction and selection in bioinformatics. It is based on data mining technique for selecting a subset of relevant features and building robust predictive models. The main idea is to choose a subset of input features by eliminating those that exhibit limited predictive performance. Feature selection can significantly improve the comprehensibility of the resulted classifier models and support the development of models that generalizes better to unseen cases.

The heterogeneity and scale of clinico-genetic data raises the demand for: (a) seamless access and integration of relevant information and data sources, and (b) availability of powerful and reliable data analysis operations, tools and services.
The challenge calls for the utilization and appropriate customization of high performing Grid-enabled infrastructures and Web technology - as presented by Web Services, and Scientific Workflows environments. Smooth harmonization of these technologies and flexible orchestration of services present a promising approach for the support of integrated genotype-to-phenotype association studies.

Grid technology. Grid computing (Foster 2003) is a general term used to describe both hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities. Grid has emerged as the response to the need for coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations. Sharing of computers, software, data, and other resources is the primary concern of Grid architectures. In a modern service oriented architecture the Grid defines the general security framework (e.g. the authentication of the users and services), the virtual organization abstraction, the user management mechanisms, authorization definition and enforcement, etc. It provides both the computational and the data storage infrastructure, which is required for the seamless management and processing of large data sets.

Semantic and Knowledge Grids. Semantic Grid presents a Grid computing approach in which information, resources and data processing services are employed with the use of semantics and respective data models. It facilitates the discovery, automated linkage and smooth harmonization of services. In a Semantic Web analogy, Semantic Grids can be defined as “extensions of current Grids in which information and services are given well-defined meaning, better enabling computers and people to work in cooperation” (De Route et al 2005). Encapsulation of Web Science and knowledge-oriented technologies in Grid-enabled infrastructures represents a flexible knowledge-driven environment referred as the Knowledge Grid (Zhuge 2004). In their layered architecture organization, Knowledge Grids define and form an additional layer, which supports implementation of higher level and distributed knowledge discovery services on a virtual interconnected environment of shared computational and data analysis resources. This setting permits and enables: automated discovery of resources; representation, creation and management of statistical and data mining processes; and composition of existing data and processing resources in ‘compound services packages’ (Cannataro and Talia 2003).

Web services. The Web Services suite of standards presents the most popular and successful integration methodology approach. Based on Web Services standards the machine-machine communication is performed via XML programmatic interfaces over web transport protocols (e.g., SOAP), which are specified using the Web Service Definition Language (WSDL) (Curbera et al 2002). These common data representation and service specification formats, when properly deployed, enable the integration of heterogeneous and geographically disparate software systems. Web Services enhance and support the development of distributed, multi-participant, and interoperable systems that can be utilized in the com-
Combination of services and their reuse as processing steps into more complex high level scenarios, commonly referred as workflows.

Scientific workflows. The Workflow Management Coalition (WFMC, www.wfmc.org) defines a workflow as “the automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules”. A workflow consists of all the steps and the orchestration of a set of activities that should be executed in order to deliver an output or achieve a larger and sophisticated goal. In essence a workflow can be abstracted as a composite service, e.g. a service that is composed by other services that are orchestrated in order to perform some higher level functionality. The (potentially parallel) steps (tasks) that a workflow follows may exhibit different degrees of complexity, and are usually connected in a non-linear way, formulating a directed acyclic graph (DAG). A Workflow Management System defines, manages and executes workflows through the execution of software that is driven by a computer representation of the workflow logic (Deelman et al 2006, Fox and Gannon 2006).

In addition to the business oriented use cases, workflows have a lot of potential in scientific areas as well. In a lot of scientific sectors, the demand is put not only on the computational power but on the complex structure of the inter-dependable tasks to be performed. Sophisticated problem-solving engages a variety of inter-dependent data analysis tasks and analytical tools, e.g., pre-processing and re-formatting of heterogeneous datasets into formats suitable as input to other analytic process. Moreover, large-scale scientific computations involve much of intervention, as in the case of the interpretation of intermediate results by domain experts. But, at some stage of the process just normal personnel could be engaged. So, the rights and roles of involved persons should be explicitly defined. In addition, the computational environment itself is heterogeneous, ranging from supercomputers to clusters of personal computers. So, there is a need to model and explicitly define the engaged computational nodes and networks. Scientific workflows are introduced as an amalgamation of scientific problem-solving and traditional workflow techniques. They have been proposed as a mechanism for coordinating processes, tools, and people for scientific problem solving purposes and aim to support “coarse-granularity, long-lived, complex, heterogeneous, scientific computations” (Singh and Vouk 1997).

To assist the bioinformatics community in building complex scientific workflows, and in the context of the EU FP6 integrated project (www.eu-acgt.org), the ACGT Workflow Editor and Enactment Environment (WEEE) have been designed and developed (Sfakianakis et al 2009). WEEE is a Web-based graphical tool that allows users to combine different Web Services into complex workflows, and it is accessible through the ACGT Portal. It supports searching and browsing of a Web Services repository and of respective data sources, as well as their orchestration and composition through an intuitive and user friendly graphical interface. Created workflows can be stored in user spaces and can be later retrieved and edited. So, new versions of them can be easily produced. Designed workflows
can be executed in a remote machine or even in a cluster of machines in the Grid. In this way there is no burden imposed on the user’s local machine since the majority of computation and data transfer of the intermediate results are take place in the Grid where the services are executed. Publication and sharing of the workflows are also supported so that the user community can exchange information and users benefit from each other’s research. WEEE is based on the BPEL (Arkin et al 2005) workflow standard and supports the BPEL representation of complex bioinformatics workflows.

The ACGT Grid environment is supported by the Gridge toolkit (www.gridge.org/) – an open source software platform, compatible with the Globus toolkit (www.globus.org) aimed to help users to deploy ready-to-use grid middleware services and create productive Grid infrastructures. All Gridge Toolkit software components have been integrated together and form a consistent distributed system following the same interface specification rules, license, and quality assurance and testing (Pukacki et al 2006).

The GG2P scenario presented in this paper is enabled by the smooth integration of components from the aforementioned technologies. GG2P aims to seamlessly integrate and mine distributed and heterogeneous clinical and genotype data sources using: (i) existing public-domain and custom-made Web Services for accessing remote and distributed genotype and phenotype data sources, and for downloading the targeted experiments and the respective data annotation (XML) files; (ii) specially devised Web Services to extract relevant information and raw data, including appropriate data pre-processing and re-formatting operations; and (iii) specially suited for G2P association studies data mining processes wrapped as Web Services. In addition, the results (profiles of specific SNPs) are automatically linked with state-of-the-art genome browsers (e.g., Ensembl), and are appropriately visualized.

3 The GG2P scenario

An SNP is a single base substitution of one nucleotide with another. With high-throughput SNP genotyping platforms massive genotyping data may be produced for individual samples (i.e., diseased, treated or, control). It is known that a category of diseases are associated to a single SNP or gene (also known as monogenic diseases). In general, a single SNP or gene is not informative because a disease may be caused by completely different modifications of alternative pathways in which each SNP makes only a small contribution. Most of the complex diseases, including cancer, are characterized by groups of genes with a number of susceptible genes interacting with each other. It’s important to search for multiple SNP profiles - among a huge number of them, that not only associate with a disease but exhibit a high discrimination power between different phenotypic classes. The GG2P scenario aims exactly towards this direction with the relevant literature started to include similar approaches (Nunkesser et al 2007, Zhou and Wang 2007,
Schwender et al 2008). The steps followed by the corresponding scientific workflow are presented and described in the sequel.

**Data access and retrieval.** Using Web Services from the European Bioinformatics Institute’s (EBI) repository (http://www.ebi.ac.uk/Tools/webservices/) we access and extract phenotypic and genotypic data from public experiments. Specifically, using specific ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae/) Web Services we may get information about a specific experiment or, get information about relevant experiments using keywords. The complete SNP array dataset used in this study is available on the NCBI GEO database under accession no. GSE3743. The dataset refers to a genotyping experiment of 78 sample hybridizations performed on the Affymetrix GeneChip Human Mapping 10K Array Xba 131 (Mapping10K_Xba131) array design. The raw data file includes 78 transformed and/or normalized data files. The hybridized samples concern breast cancer (BRCA) and normal (CTRL) cases. More information about the dataset can be found at (Richardson et al 2006). Note that GG2P could be easily customized to work with other experiments and respective datasets.

**Data mediation.** The response of ArrayExpress web service is an XML file with links to phenotypic (via the ‘sdrf’ tag) and genotype (via the ‘fgem’ or ‘raw’ tags) experimental data (see Fig 3.1 for a sample of the XML response file). We utilized a special parser to extract the needed information from the XML file.

```xml
<experiment total-assays="78" total-samples="78" total="1" revision="080925" version="1.1">
<experiment>
<id>1627324147</id>
<accession>E-GEOD-3743</accession>
<name>Genotyping of human breast tumors</name>
<samples>78</samples>
...
<files>
<raw celcount="78" count="78" name="E-GEOD-3743.raw.zip"/>
<fgem count="78" name="E-GEOD-3743.processed.zip"/>
<idf name="E-GEOD-3743.idf.txt"/>
<sdrf name="E-GEOD-3743.sdrf.txt"/>
<biosamples>
<png name="E-GEOD-3743.biosamples.png"/>
<svg name="E-GEOD-3743.biosamples.svg"/>
<biosamples>
<files>
</experiment>
</experiments>
```

**Fig. 3.1.** Part of Web Service XML response file (from ArrayExpress)
The parser locates the ‘samples’, ‘sdrf’ and ‘fgem’ tags. The ‘samples’ tag identifies the number of included samples/hybridizations, and the ‘sdrf’ tag points to the respective file with description of each hybridization. From the ‘fgem’ tag we may identify and download the SNP profiles of the respective experiment’s samples. It is essential to align phenotypic classes with the respective samples’/hybridizations’ genotype data, and form a unified dataset to be analyzed. We employ a natural-language mechanism, enabled by specific ontologies and controlled vocabularies (Potamias et al 2005). The result is a homogenized and appropriately formatted file (with phenotype class annotations and respective geno-type data), which serves as input to a specific analytical process.

**Data preprocessing.** Depending on the data and the data mining algorithm, the formed data file may need extra processing. For example, many algorithms can handle only nominal values. In such a case, and if the data comes with continuous feature values, we have to discretize them. Furthermore, as genotype profiling platforms (like Affymetrix) produce too many ‘NoCalls’, one may be also interested to reduce these ‘missing values’ utilizing an appropriate data pre-processing process. After the needed pre-processing are performed, the ‘filtered’ dataset is transformed into the ARFF format - a de facto standard for machine learning. ARFF supported by the Weka machine learning package (http://www.cs.waikato.ac.nz/ml/weka/) (Witten and Frank 2005).

**Data analysis.** A variety of existing data mining algorithms exists in the public domain (e.g., Weka, R-package/Bioconductor, BioMoby). Here we rely on a feature reduction and selection approach. Dimensionality reduction and feature selection is a well-known and addressed issue in machine learning and data mining (Guyon and Elisseeff 2003). We are interested on the identification of SNP-phenotypic class associations, and on respective discrimination/classification models. The profiles of these SNPs are able to distinguish between particular pre-classified patient samples. Core operations of this process are implemented in the MineGene gene selection system, and their Web Services deployment (Potamias et al 2004, Potamias et al 2006).

### 3.1 GG2P in action

For the realization of GG2P scenario we used part of the ACGT Grid infrastructure – the Data Management System, the service repository and the workflow editing and execution environment. The Data Management System (DMS) is a secured and distributed file system over the Grid. The service repository gives access rights as well as metadata information about the available services. The workflow editor is a Web2 application and, as already mentioned, the workflow enactor is a BPEL-compliant application installed in a Grid node. Fig. 3.2 introduces the GG2P knowledge discovery scenario as implemented in the context of the ACGT WEEE workflow editing and execution environment. The Web Ser-
vices (not shaded shapes in the workflow area of Fig. 3.2) are registered in the ACGT services repository.

The ACGT environment requires authorization from the DMS and the services repository. DMS grants permissions to user’s account in the Grid and services repository give access to available services. Then the user composes and draws the desired workflow. At the next step, the editor translates (or compiles) the graphical workflow into BPEL. Finally, the enactment of the workflow may start. The first web service takes as input a query (first, from left, shaded shape of Fig. 3.2) and returns an XML file with information about all the related to the query experiments in the EBI ArrayExpress repository. For the specific scenario we used a query with the keywords “homo sapiens” & “breast cancer” & “genotype” & “affymetrix” & “Mapping10K_Xba131”.

Fig. 3.2. The GG2P scientific workflow as implemented in ACGT’s Workflow Editor and Enactment Environment (WEEE). Web services include: ArrayExpress, Mediator, Discretization, and Data Mining. Services are activated by a Query (top part). Deployment of Data Mining also needs specification of parameters (‘Param 1’ and ‘Param 2’)

The second service (Mediator) takes as input the repository’s XML response file and creates the homogenized file with the clinical and genotype data. The generated file is stored in DMS at the user’s account. The next service (Discretization) discretizes and transforms the experiment data to arff format. Discretization service retrieves the data from DMS and stores the arff-formatted data back to the DMS. The final service implements the (two-valued) SNP feature selection algorithm. The service again retrieves data from DMS and stores the results in the DMS. Then, after the editor requests the results from the DMS, SNP annotations and links to the Ensembl genome browser are automatically assigned to the se-
lected SNPs. Finally, an html file is formed and is used for the visualization of results (see Fig. 4.1).

4 Results and Discussion

The Affymetrix SNP genotyping platforms produce processed data files where, each SNP receives three different values: AA and BB that represent paternal or maternal homozygosity statuses, respectively, and AB for heterozygosity ones. The ‘0’ and ‘1’ nominal values are assigned to the AA/BB and AB SNP feature values, respectively. This results into a two-valued feature representation space. In this setting a set of SNPs could be considered as an ideal discriminator between two different phenotypic classes if it displays the ‘0’ value for all sample cases in one class and the ‘1’ value for all sample cases in the other class. From the total of the 78 sample cases included in the target SNP genotyping experiment we excluded the ones that have more than 10% of missing ‘NoCall’ values, resulting into a dataset of 36 BRCA and 36 CTRL cases.

For the target BRCA vs. CTRL study, the execution of the GG2P scientific workflow resulted into a set of about 100 most discriminant SNPs. With these SNPs the following highly performing figures are achieved: 96.2% accuracy, 92.2% sensitivity, 96.2% specificity, and 0.979 ROC/AUC.

Fig. 4.1 visualizes just the top 24 of them with the highest ranks (for those sample cases with no ‘NoCall’ SNP values) sorted by their chromosomal location. The first column shows the discrimination power (the rank) for each SNP (as calculated by MineGenes’ core feature selection process). The second column shows the Affymetrix code name for the probe that represents the respective SNP. The third column displays the corresponding code, namely: dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/). The dbSNP - SNP databases, represent a widely used public-domain archive for a broad collection of SNPs as well as small genomic insertion/deletions (indels) and is hosted at the National Center for Biotechnology Information (NCBI). The next three columns display information about the genomic region of the respective SNP: column four the chromosomal location; column five the cytoband, and columns five and six the nucleotide allele variations for the two (paternal/maternal) alleles. The last column shows the nearest gene present in the corresponding SNP’s genomic physical position.

All hyperlinks are automatically assigned to the respective items by consulting the annotation files provided by Affymetrix. When clicking on a specific cytoband one is transferred to the respective visualization screen of the Ensembl genome browser (www.ensembl.org). So, inspection of results and further investigation is enabled and supported. In Fig 4.1 one may also observe and contrast the SNP characteristic profile patterns between BRCA and CTRL cases, respectively - gray and dark shaded cells represent homozygosity (‘AA/BB’) and heterozygosity (‘AB’) statuses, respectively.
The main observation is that the homozygosity patterns are dominant in the BRCA cases - a finding which is consistent with the **Loss of Heterozygosity** (LOH) situation in pathogenic situations. LOH in a cell represents the loss of regular function of one of the gene’s alleles when the other allele is inactive. In oncology, LOH refers to somatic mutations and occurs when the offspring’s functional allele is inactivated by the mutation. In such situations, normal tumor suppressor functionality is inactivated and tumorigenesis events are almost certain.

Fig. 4.1. The induced most discriminant and highest ranked BRCA vs. CTRL SNPs (for the ArrayExpress E-GEOD-3743 genotyping experiment) – gray shaded and dark shaded cells indicate homozygosity and heterozygocity statuses, respectively. It can be easily observed that LOH (Loss Of Heterozygosity) patterns dominate the BRCA cases.

We further examined the biological relevance of the findings, i.e., does the identified and most discriminant SNPs relate to LOH and breast cancer situations. Literature search provide us with strong evidence for that. We refer to just two indicative SNPs in cytobands 17p13.2 and 17p12 (both highly ranked). Chromosome 17p is among the most frequently deleted regions in a variety of human malignancies including breast cancer. In (Seitz et al 2001) the localization of a putative tumour suppressor gene (TSG) at 17p13, distal to the TP53 (the most indicative tumor suppressor) gene, was further refined for breast carcinomas. It was found that 73% (37 of 51) of the breast tumors exhibited loss of heterozygosity (LOH) at one or more loci at 17p13. The allelic loss patterns of these tumours suggest the presence of at least seven commonly deleted regions on 17p13. The three most frequently deleted regions were mapped at chromosomal location 17p13.3 - 17p13.2. Furthermore, the data suggest that different subsets of LOH in this region are associated with more aggressive tumor behavior. Additional evidence for the association between the 17p13 genomic region and breast cancer are also reported in (Mao et al 2005) and (Ellsworth 2003). Similar findings are re-
ported for the 17p12 region. In (Shen et al 2000) sixty-three markers are reported that display ≥25% LOH, with the highest values being observed on 17p12 (48.4% for the well, and ~87% for the poorly differentiated breast tumor cases).

5 Conclusions and Future Work

We presented an integrated methodology that enables the discovery of genotype-to-phenotype associations and predictive models, and supports G2P association studies. The methodology is realized in the context of the GG2P scenario being implemented with the aid of Web Services and Scientific Workflows and operating in a grid environment. In particular the ACGT (EU FP6 integrated project) Grid infrastructure and its WEEE workflow editing and enactment environment were utilized.

The GG2P workflow was executed on an indicative SNP genotyping experiment (from the ArrayExpress repository) that concerns the hybridization breast cancer and normal/control tissue samples. We were able to identify about 100 indicative SNPs that exhibit contrasted homozygosity / heterozygosity profiles, and achieve highly discriminant performance figures for the respective phenotypic classes. The most highly ranked SNPs exhibit clear loss of heterozygosity patterns, a common situation in tumorigenesis. Literature searches provide strong evidence about the biological relevance of the findings – the respective SNP’s genomic regions are strongly association with characteristic breast cancer phenotypes.

Our immediate R&D plans, among other, include: experimentation with other public-domain genotyping experiments, and enrichment of GG2P and its workflow realization with other data-mining techniques (e.g., clustering, association rules mining etc).

Acknowledgements: This work is partially supported by the European Commission’s Sixth and Seventh Framework Programme in the context of the ACGT (FP6-2005-IP-026996) and GEN2PHEN (FP7 HEALTH-F4-2007-200754) Integrated projects, respectively.

References

Deelman E, Zhao Z, Belloum A (Eds.) (2006) Scientific Programming Journal, special issue on workflows to support large-scale science 14(3–4)
Cross-platform integration of transcriptomics data

Georgia Tsiliki, Marina Ioannou and Dimitris Kafetzopoulos

Abstract An increasing number of studies have profiled gene expressions in tumor specimens using distinct microarray platforms and analysis techniques. With the accumulating amount of microarray data, one of the most challenging tasks is to develop robust statistical models to integrate the findings. This article reviews some recent studies on the field. We also study the intensity similarities between data sets derived from various platforms, after appropriate rescaling of the measurements. We found that intensity and fold-change variability similarities between different platform measurements can assist the analysis of independent data sets and can produce comparable results with those obtained for the independent data set alone.

1 Introduction

With the increasing availability of published microarray data sets there is a need to develop approaches for validating and integrating results across multiple studies. The overlap of gene expression signatures of various studies is very small, for example between the “Amsterdam” signature [23] and the “Rotterdam” signature [24], mainly due to the small sample sizes of individual studies and error measurements. A major concern in the “meta-analysis” of DNA microarrays is the lack of a single standard experimental platform for data generation. The microarray technologies

Georgia Tsiliki
FORTH, Institute of Molecular Biology & Biotechnology, P.O. Box 1385, 711 10, Heraklion, Greece, e-mail: tsiliki@imbb.forth.gr

Marina Ioannou
FORTH, Institute of Molecular Biology & Biotechnology, P.O. Box 1385, 711 10, Heraklion, Greece, e-mail: mioannou@imbb.forth.gr

Dimitris Kafetzopoulos
FORTH, Institute of Molecular Biology & Biotechnology, P.O. Box 1385, 711 10, Heraklion, Greece, e-mail: kafetzo@imbb.forth.gr
currently in use differ in how DNA sequences are laid on the array, the length of these sequences, splicing variations and the number of samples measured in each hybridization. As a result, an important source of technological variability in gene expression measurements is the platform used.

The increasing number and availability of large-scale gene expression studies of human and other organisms provide strong motivation for cross-study analyses that combine existing and/or new data sets. In a cross-study analysis, the data, relevant test statistics or conclusions of several studies are combined. Several studies have compared measurements across platforms [9] and reported their findings in terms of reproducibility of results, power increase of studies, validation of gene signature results [10] [8] [11] [25]. The MAQC Quality Control Consortium, the FDAs Critical Path Initiative, NCI's caBIG and others are implementing procedures that will broadly enhance data quality. The MAQC consortium have reported that proper sample preparation is sufficient to dramatically enhance multi-lab and multi-platform correlations [16].

However, combining data from different expression studies and possibly different gene expression platforms poses a number of statistical difficulties due to the different processing facilities. As a consequence, measurements from different platforms cannot be directly combined. Identifying and removing such systematic effects is the primary statistical challenge in cross-study analysis. We note that technological differences between studies may be confounded with biological differences arising from the choice of patient cohorts (e.g. age, gender or ethnicity). In many cases, technological artifacts are dominant, though care should be taken to verify this, and one can hope to remove them while leaving biological information intact.

Here we briefly review some recent techniques to minimize error measurements and safely combine results of studies which address the same biological questions. Furthermore, we evaluate how the direct use of intensity data from independent data sets and platforms, can facilitate the statistical analysis of other microarray studies. An advantage of such an approach is that the same methodology can be used and the measurement errors can be controlled in the same way for all data sets. Our scope is to demonstrate that the power of any statistical conclusions can be retained when the data is enhanced with external data from various platforms. For that purpose our working example is the classification of ER samples in a breast cancer data set.

1.1 Recent literature review

In general, it will make sense to combine data sets of studies which address the same questions, or, experiments with some sufficiently similar aspects so that one can hope to make better inference from the whole than from the experiment separately. However, in order to compare experiments that are performed on different gene expression platforms, the first thing one should look at is how to link oligonucleotide probe sets, spotted sequences, and other microarray features. Typically, a sequence-specific identifier (GenBank accession number) serves as a reference to the array
probe sequences. Thus, the first step in a cross-study analysis would be to identify a subset of genes which are consistently measured across platforms. The next step would be to derive for each individual data set numerically comparable quantities from the expression values of genes in the common list by applying specific data transformation and normalization methods.

The most simple approach to integrate data would be to sample standardize and gene median center each available data set, and then combine data sets. More systematic approaches have been proposed for integration of findings from multiple studies using different array technologies. Particularly, according to [14], there are several potential approaches to cross-study analysis, depending on what information is being synthesized. Existing studies either combine information from primary statistics (such as t-statistics or p-values) [13] [19] or secondary statistics (such as gene lists) that are derived from the individual studies [3]. Additionally, other approaches to meta-analysis of gene-expression data are considered by [4] [15] [12], which directly integrate the data and then proceed with the analysis.

[22] proposed optimization methods for cross-laboratory and cross-platform microarray expression data, based on three simple and often employed techniques to identify discrepancy in expression data sets. They created an experimental design that compared three functionally different normal tissues: human liver, lung and spleen. Particularly, they reported that when precision, biological interpretation and multiple platform data sets were considered together, they allowed for better selection of genes with respect to a particular outcome. They considered precision and sensitivity measurements which were useful in finding the minimal detectable fold-change and raw performance values for an array platform. Also, Gene Ontology and pathway analyses were considered, which were thought to be a valuable way of examining and comparing the actual biological interpretation. Differences in pathways indicated consistency problems which could be quantified by counting the differentially expressed genes between platforms that moved in different directions.

Along these lines, [25] integrated three independent microarray gene expression data sets for breast cancer and identified a structured prognostic signature consisting of 112 genes organized into 80 pair-wise expression comparisons. The method used for integration of data sets was based on the ranks of the expression values within each sample first introduced in [5]. Since the features were rank-based, data normalization was not necessary before data integration.

A cross-study normalization method called $XPN$ was suggested by [14], which based on identifying homogeneous groups of genes and samples in the combined data. Specifically, they employed k-means clustering independently to genes and samples of the combined data to identify blocks (or clusters) in the data. Then, each gene expression value was a scaled and shifted block mean plus noise. Their model assumed that the samples of each available study fall roughly into one of the statistically homogenous sample groups identified, and that each group was defined by an associated gene profile that was constant within each of the estimated gene groups. They examined three existing breast cancer data sets and reported that XPN successfully preserved biological information according to ER prediction error rates while removing systematic differences between platforms.
The reliability of gene expression across three previously published breast cancer studies was evaluated by [4]. They compared the strength of evidence of gene to phenotype associations across studies and combined effects across studies. Their methods are implemented by [2] on an R package (www.r-project.org) library called MergeMaid (http://www.bioconductor.org/packages/2.2/bioc/html/MergeMaid.html). They defined a reliability score and set a threshold via permutations to distinguish which were the “reliable” genes in two study experiments, i.e. the genes consistently measured in all studies. For multi-study experiments they considered an alternative interclass correlation coefficient per gene. Finally, they used a between studies combined effect based on the first eigenvector of a principal component analysis (PCA) of each study, to determine the genes that are associated with the phenotype.

In order to account for inter-study variation, [3] suggested an “effect size” model for multiple microarray studies. They defined effect sizes as standardized indexes measuring the magnitude of a treatment or covariate effect. They suggested the use of a fixed-effects model (FEM) or a random-effects model (REM) (or alternatively a hierarchical Bayesian model) depending on the homogeneity of study effects. Finally, they measured the statistical significance of their combined results by permutation tests and FDR calculations. Many of their methods are implemented in GeneMeta R package library (http://www.bioconductor.org/packages/2.2/bioc/html/GeneMeta.html).

Finally, an interesting approach is that by [15] who applied a two-stage Bayesian mixture modeling strategy to analyze four independent breast cancer microarray studies derived from different microarray platforms (spotted cDNAs, Affymetrix GeneChip, and inkjet oligonucleotides). They derived an inter-study validated 90-gene “meta-signature” predictive of breast cancer recurrence. Their analysis was based on the signed conditional probabilities of differential expression as introduced in [12]. Particularly, [12] proposed a Bayesian mixture model transformation of DNA microarray data with potential features applicable to meta-analysis of microarray studies, although they employed them in the context of molecular classification. The basic idea was to estimate the platform independent probability of over-expression, under-expression or baseline expression for gene sample combinations given the observed expression measurements. Along these lines, [15] reported that the use of the specific probability measures increased the power of statistical analysis by increasing the sample size.

There is a great challenge to compare and integrate results across independent microarray studies. Meta-analysis studies sometimes produce comparable results even under different logics. Although all approaches, normalization or combination of secondary results, have their merits, here we proceed with studying the effects of scaling existing measurements from various platforms as that was suggested by [12]. An important selection criterion for data integration is the measurement correlations between platforms [18]. Nonetheless, a large number of genes might be lost when looking at the correlation due to different levels of noise between platforms. We find that rescaling of measurements should be able to prevent that.
2 Integrate findings

Here we suggest some characteristics of the data that need to be accounted for when assimilating results from different studies, and evaluate them in independent data sets. Particularly, we consider the “translation” procedure of values as that was first suggested by [12] and it was employed by [15] on the same content. They estimated probabilities of over-expression, under-expression and baseline expression, and translated the intensity measurements into a probability of differential expression. The new probability scale can make comparisons between platforms on a unified scale rather than using gene-specific summaries. For an analytic description of the method see [15].

We use the four data sets also considered by [15], namely the [20], [21], the [23], [7] data sets. The first two studies are cDNA microarray studies, the third is an in situ oligonucleotide array study and the fourth an Affymetrix GeneChip study. The data sets consist of 305 breast cancer samples in total and 2,555 common genes. The study-specific breast cancer prognosis signatures have been previously reported to have a small overlap. [15] suggested that combination of the four in a probability scale derives a 90 gene meta-signature which is strongly associated with survival in breast cancer patients. We study their approach in terms of the sample’s ER status categorization. Furthermore, we suggest a few modifications which seem to strengthen our results in an independent data set produced with homemade two-colour spotted arrays from Qiagen V3 human library. All results presented here are with respect to that independent data set which consists of 34,772 70mer probes and 29 samples (18 ER+ and 11 ER− samples). We refer to that data set as Data1 from here onwards.

Measurements for all four data sets [20] [21] [23] [7] considered here are on the so called “poe” scale [15] and vary in the interval [0, 1]. Our scope is to measure the accuracy of sample classification with respect to their ER status by using simple statistical measures. For that reason, we only consider t-test calculations and Ward’s hierarchical clustering with euclidean distance. We avoid comparing our results with those derived when studies are considered individually, since those finding are based on a more advanced statistical methodology. Thus, our scope is to compare the ER classification outcome in Data1 samples when it is assisted by external data and under the same statistical methodology.

2.1 ER signatures when combining data sets

If we consider all 304 samples (one sample from [23] data set had an unknown ER status and was excluded from further analysis), we find a set of 272 genes adequate to distinguish the two classes (ER+, ER−). From those we found 75 common with Data1. There are some common genes with those reported by [23], for example, for ER categorization. Particularly, [23] reported a set of 550 genes, from which 223 are common with Data1. However, only 12 genes are common between the
two list and can be found in Data1. In Figure 1 we can see the two ER signatures. An interesting observation is that both appear to have two mis-classification errors.

We apply agglomerative hierarchical clustering algorithm using Euclidean distance metric and Ward clustering algorithm [13].

Alternatively, if we consider the whole of Data1 and apply the same methodology as before, we find 279 genes able to statistically distinguish ER. We refer to those results as the Intrinsic Model results. However, it would be interesting to consider only the genes of Data1 which are common with the [20] [21] [23] [7] data sets. In this case, 120 statistically significant genes are able to distinguish the two ER classes. Those results are refer to as the Starting Model results.

### 2.2 Intensity and fold-change similarities

Many times the intensity measurements vary between platforms for their common probes. That variability could indicate platform specific effects, or even random noise due to experiment conditions. In this subsection, we study how that variability can affect an ER derived signature which is based on many platforms. For that reason, we consider only probes that appear to have “similar” values across the four data sets in terms of magnitude on the “poe” scale. Particularly, since we are interested on ER classification, we search for genes with similar intensity behaviour in separately ER+ and ER− samples.

We employ Kruskal-Wallis rank sum tests [6, p.115] per gene, to test the null hypothesis that the location parameters of the distribution of ER+ and ER− samples
are the same in each of the four data sets. The alternative is that they differ in at least one. We consider only genes with high p-values for both ER+ and ER− samples, which based on the test give evidence for accepting the null. The left hand-side plot of Figure 2 shows the 44 genes that appear to have the same location distribution parameters for both ER+ and ER− samples across the four data sets. For the right hand-side plot we consider 100 permutations per gene and finally report only 65 genes with significant permutation based empirical p-values with respect to ER status. We can observe that the mis-classification errors are three in both cases, however, permutation procedure is inferior in terms of the number of genes included.

Another characteristic of the data is the fold-change behaviour between the ER+ and ER− samples. When we consider genes with the same amount of fold-change variability across the four data sets, we find that 24 genes, common for the four data sets and Data1, could distinguish the two ER classes. The genes were selected to have the same fold-change levels for the four platform measurements examined here. In Figure 3 we can see that the two ER classes can be well distinguished and in this case.

2.3 Results

In order to evaluate the approaches suggested before and account for statistical sampling error, we employ multiclass bootstrap resampling techniques and estimate via probabilistic measures whether clusters of the original data found by hierarchical
clustering are strongly supported by the data. For that reason we calculate two types of p-values as they are defined in [17]; the *approximately unbiased* (AU) p-value and the *bootstrap probability* (BP) value. AU p-value is computed by multiscale bootstrap resampling and is thought to be a better approximation to unbiased p-value than BP value which is computed by normal bootstrap resampling. However, the AU p-values themselves include sampling error, since they are also computed by a limited number of bootstrap samples. The null hypothesis in this case is that the clusters of the data are observed by chance. Clusters with AU p-values higher than 95% are strongly supported by data, i.e. those clusters do not seem to be caused by sampling error, but may stably be observed if we increase the number of observations.

[17] suggested that 10 sample sizes for each data set should be examined. Along these lines, we consider sample sizes equal to the $r' = \{0.49, 0.6, 0.69, 0.8, 0.89, 1.0, 1.09, 1.2, 1.29, 1.4\}$ percentages of the original sample size. For each sample size we generate 10,000 bootstrap samples. For each bootstrap sample, we apply hierarchical clustering to obtain the sets of bootstrap replications of dendrograms and compute the BP for observing each cluster. Finally, we estimate AU p-values by fitting a regression model to the BP values calculated for each cluster and each sample. For an analytic description of the method see [17].

In Table 1 we report the AU and BP values for the approaches already mentioned for the two major clusters of the data $C_0$ and $C_1$, where $C_0$ mostly contains ER− samples and $C_1$ mostly contains ER+ samples. We also report the frequency of misclassified samples in $C_0$ and $C_1$, and the number of statistically significant genes with respect to ER status. Note the decrease in the number of significant genes because of mapping when information from combined data is used. The results in the first
row of the table (Intrinsic Model) correspond to clustering results after t-test calculations are directly employed to Data\textsubscript{1}, whereas, results in the second row (Starting Model) refer to Data\textsubscript{1} but only to its common genes with the four data sets. We consider those values as a baseline for comparison with other approaches suggested here. The results in the third row (Simple Model) correspond to clustering results derived from the four merged data sets. Particularly, we found the significant genes, with respect to the ER status, when the four data sets were considered together and after Benjamini-Hochberg correction was applied, and applied our finding to Data\textsubscript{1}. The Fold-change variability, Kruskal-Wallis (K-W), K-W with Permutations results correspond to methods presented in section 2.2.

Table 1 We report the AU and BP values from bootstrapping, the frequency of mis-classified samples and the number of statistically significant genes with respect to ER status. For each variable the two values corresponds to clusters C\textsubscript{0} and C\textsubscript{1}, respectively. K-W corresponds to Kruskal-Wallis method. Results are reported for Data\textsubscript{1}.

<table>
<thead>
<tr>
<th>Approach</th>
<th>AU (%)</th>
<th>BP (%)</th>
<th>Mis-classifications Freq.</th>
<th>Num. Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic Model</td>
<td>88 - 83</td>
<td>69 - 52</td>
<td>0 - 0.182</td>
<td>279</td>
</tr>
<tr>
<td>Starting Model</td>
<td>89 - 88</td>
<td>38 - 34</td>
<td>0 - 0.182</td>
<td>120</td>
</tr>
<tr>
<td>Simple Model</td>
<td>75 - 79</td>
<td>10 - 8</td>
<td>0.111 - 0.15</td>
<td>75</td>
</tr>
<tr>
<td>Fold-change Variability</td>
<td>93 - 95</td>
<td>34 - 34</td>
<td>0 - 0.25</td>
<td>24</td>
</tr>
<tr>
<td>K-W</td>
<td>76 - 79</td>
<td>27 - 10</td>
<td>0 - 0.182</td>
<td>44</td>
</tr>
<tr>
<td>K-W with Permutations</td>
<td>86 - 78</td>
<td>12 - 7</td>
<td>0 - 0.182</td>
<td>65</td>
</tr>
</tbody>
</table>

We can observe that the K-W and Simple Model results have similar AU p-values, although the number of significant genes is higher in the second case. However, they both have smaller AU p-values compared to the Starting Model. Better results in terms of AU p-values and number of genes, can be observed in the case of permutation sampling with Kruskal-Wallis tests. The number of genes increases from 44 to 65 and the AU p-values are elevated supporting the alternative hypothesis that C\textsubscript{0} and C\textsubscript{1} clusters are not observed by chance. However, Fold-change Variability results exhibit the highest AU p-values, although the number of significant genes is small compared to that of the other approaches. The mis-classification frequency is relatively small in all cases, whereas the BP values are variable compared to the AU.

To prove the power of a high number of independent data sets used, in Table 2 we focus on the fold-change variability results but for only three data sets ([23] [20] [21]) and two data sets ([23] [21]) chosen at random from the four. We can observe that our results benefit in terms of AU p-values when information from more data sets is used.
Table 2 We report the AU and BP values from bootstrapping, the frequency of mis-classified samples and the number of statistically significant genes with respect to ER status. For each variable the two values corresponds to clusters \( C_0 \) and \( C_1 \), respectively. K-W corresponds to Kruskal-Wallis method. Results are reported for Data1.

<table>
<thead>
<tr>
<th>Approach</th>
<th>AU (%)</th>
<th>BP (%)</th>
<th>Mis-classifications Freq.</th>
<th>Num. Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold-change Variability</td>
<td>93 - 95</td>
<td>34 - 34</td>
<td>0 - 0.25</td>
<td>24</td>
</tr>
<tr>
<td>F-c V 3 data sets</td>
<td>70 - 75</td>
<td>25 - 20</td>
<td>0 - 0.182</td>
<td>29</td>
</tr>
<tr>
<td>F-c V 2 data sets</td>
<td>63 - 69</td>
<td>20 - 19</td>
<td>0 - 0.143</td>
<td>31</td>
</tr>
</tbody>
</table>

3 Conclusions

We considered how information from studies using various platforms can facilitate the search for significant genes with respect to the categorization of ER samples. Our analysis focused on ER status classification although other parameters, binary or continuous such as breast cancer prognosis, could be studied. An obvious limitation of such approaches is the restriction of the study to only annotated common probes.

We studied the effect of rescaling measurements from four platforms to a common scale and use the information obtained by that data. We employed resampling techniques to minimize sampling error and variability introduced by the different platforms. Our results were compared to those obtained from direct analysis of data, and were thought to be able to describe properties of independent data sets. Particularly, we found that an important property in such kind of analyses is the fold-change variability of common probes across various studies. The performance of K-W analysis was also comparable to that of direct analysis, when data was enhanced with permutations. In all cases, gain in terms of AU p-values resulted in loss of some genes. Overall, we showed that knowledge from numerous data sets produced under the same biological question, can greatly assist the statistical analysis of independent data sets.

References


Method for relating inter-patient gene copy numbers variations with gene expression via gene influence networks

Sylvain Blachon, Gautier Stoll, Carito Guziolowski, Andrei Zinovyev, Emmanuel Barillot, Anne Siegel and Ovidiu Radulescu

Abstract During tumorigenesis, genetic aberrations arise and may deeply affect the tumoral cell physiology. It has been partially demonstrated that an increase of genes copy numbers induces higher expression; but this effect is less clear for small genetic modifications. To study it, we propose a systems biology approach that enables the integration of CGH and expression data together with an influence graph derived from biological knowledge. This work is based on 3 key ideas. 1) Inter-individual variations in gene copy number and in expression allow to attack tumor variability and ultimately addresses the problem of individual-centered therapeutics. 2) Confronting post-genomic data to known regulations is a good way to check the soundness and limits of current knowledge. 3) The abstraction level of qualitative modeling allows integration of heterogeneous data sources. We tested this approach on Ewing tumor data. It allowed the definition of new biological hypotheses that were assessed by random permutation of the initial data sets.

Sylvain Blachon
IRISA, UMR 6074 CNRS/INRIA/Universite Rennes 1, campus de Beaulieu, F, 35042 Rennes Cedex, France e-mail: sylvain.blachon@irisa.fr

Gautier Stoll,
UMR 900 INSERM/Mines ParisTech/Institut Curie, 26 rue d’Ulm 75248 PARIS cedex 05, France

Carito Guziolowski
IRISA

Andrei Zinovyev
UMR 900 INSERM/Mines ParisTech/Institut Curie

Emmanuel Barillot
UMR 900 INSERM/Mines ParisTech/Institut Curie

Anne Siegel
IRISA

Ovidiu Radulescu
IRMAR, UMR 6625 CNRS/Universite Rennes 1, Campus de Beaulieu, 35042 Rennes cedex, France
Abreviations : GCNv = Gene Copy Number Variation ; ELv = Expression Level Variation; ES = Ewing Sarcoma

1 Introduction

Relating genomic instabilities to gene expression is a difficult challenge which is not yet completely resolved. The biological hypothesis is that a gene amplified genomically (in tumor cells for example) induces a higher expression.

Relating gene expression profiles to gene copy numbers was mostly performed using correlation analyzes, in order to find candidate genes serving as markers or as potential targets for therapy [12].

However, these correlation analyzes cannot explain all gene behaviours : in the best case, 50% of them can be explained [7, 10]. This proportion is much weaker for tumors that have less instabilities (like Ewing sarcoma) than more common tumors, like breast cancers. Hence, on those tumors, it appears difficult to extract relevant global properties to relate CGH data to tumor outcomes [8, 16, 11, 1] or to gene expression [3, 15].

We proposed new method for the study of genomic instabilities in tumors, based on the systems biology approach. In this approach, we include the biological processes that regulate transcription through the dynamics of one or several networks of interacting molecules. In such a model genes, transcripts and proteins are network components. The simple process one-gene-one-transcript-one-protein is replaced by a more global point of view involving all the connections among the network components.

In order to deal with small genetic modification, we adopt a more mechanistic approach to genetic variability via a network model. Genetic variability, having nowadays interesting perspectives in personalized medicine, has been addressed by various biologists since Darwin. The idea that interaction between genes can modulate the effects of this variability can be traced back to Conrad Waddington, whose chreods can be interpreted as representations of the “elastic” response of gene networks. Here, gene-gene interactions can stabilize the effect of genetic variability. However, this can be done only up to a certain extent, as some variability is necessarily persistent. The persistent variation is not entirely random, it bares information on the network.

Based on these ideas, a framework was conceived to address the following questions :

• how are gene copy number (GCN) and expression level (EL) variations related?
• is a (theoretical) gene regulation model consistent with (real life) observed variations in patient pairs? If so, what is the GCN contribution to consistency?

This framework is based on qualitative reasoning which formalizes biological interactions [13] and is efficient ¹ on large scale regulatory networks [5, 17].

¹ see also the web interface : http://www.irisa.fr/symbiose/bioquali
We applied our methods to Ewing sarcoma: it is a pediatric bone tumors that originate from a translocation t(11;22)(q24;q12), producing a chimeric gene: \textit{EWS-FLI1} [2]. This chimeric gene is thought to act as an aberrant transcription factor. A set of target genes that can be either activated or repressed has been already discovered [14].

2 Available data and model for Ewing sarcoma

2.1 Data description and preprocessing

A home made Comparative Genomic Hybridization (CGH) array was built in-house at Institut Curie (3920 probes 60 bp long covering all the chromosomes). CGH data were produced on a set of 47 tumors, including 7 cell lines.

Among the 39 remaining tumors, 12 were diagnosed as metastatic tumors before analysis and therapy. After analysis and therapy, on the 27 remaining tumors, 10 evolved in metastasis while 17 remained localized tumors.

An Affymetrix U133A chip was used to measure expression levels on the biopsies of patient tumors. Microarray data were normalized by the GC-RMA technique.

Breakpoint detection on CGH data was performed using GLAD algorithm[6]. GLAD allows CGH level smoothing in a given genomic region flanked by two breakpoints.

2.2 Model description

A gene regulation model involving 130 genes, including \textit{EWS−FLI1}, was designed within SITCON project [4]. The genes/pathways included in this network model were indentified by analysis of transcriptome time series on Ewing cell lines. The logical connections between genes are based on 1) scientific literature and 2) manually curation of TRANSPATH [9] database. Main tumor phentypes are included in this network: cell cycle regulation, apoptosis and cell migration.

3 Systems biology method for analyzing genetic and expression variations

Our methodology aims at confronting pairwise variations with the \textit{(EWS−FLI1)} gene network model described above.

\footnote{One was too noisy and was discarded from the analysis}
In order to cope with genetic variability in gene networks, we represent differences between individuals as perturbations of the network. Biologically, the hypothesis is that data obtained on a cell population coming from a tumor biopsy reflect a molecular steady state. For a patient pair, the whole set of observed variations describes the qualitative differences between the two steady states. This can be coherently done in a framework that was first introduced in [13], based on interaction graphs and qualitative equations.

### 3.1 Qualitative equations

Consider a network of \( n \) interacting components. The interaction model is the digraph \( G = (V,E) \), \( V = \{1,\ldots,n\} \). There is an edge \( j \rightarrow i \in E \) if \( j \) influences the production of \( i \). Edges are labelled by a sign \( \{+,-\} \) which indicates whether \( j \) activates or represses the production of \( i \). Let us denote by \( \text{sign}(\delta X_i) \in \{+,-,?\} \) the sign of the variation of \( i \) between two conditions, and by \( \text{sign}(j \rightarrow i) \in \{+,-\} \) the sign of the edge \( j \rightarrow i \) in the interaction graph.

For every predecessor \( j \) of \( i \), \( \text{sign}(j \rightarrow i) \cdot \text{sign}(X_j) \) provides the sign of the influence variation of \( j \) on the species \( i \). Notice that this can be either positive (increased activation or decreased repression) or negative (decreased activation or increased repression). Then, the constraints that the network imposes on the variations can be expressed as qualitative equations:

\[
\text{sign}(\delta X_i) \approx \sum_{j \rightarrow i} \text{sign}(j \rightarrow i) \text{sign}(\delta X_j). \tag{1}
\]

The sign algebra is summarized in the following table.

\[
\begin{array}{cccc}
++ &=& + & \times \times = & + \\
\end{array}
\]

### 3.2 Taking into account genetic variations

In order to take into account the genetic variability of the patients we introduced new qualitative variables representing, for a given pair of patients, the GCN variations. The corresponding nodes in the interaction digraph will be called “gene nodes”. There is one gene node for each gene considered and in our analysis we kept a set of 126 genes. The remaining nodes are either mRNA or protein nodes occurring in the \((EWS − FLI1)\) network.
The central hypothesis here is that gene nodes act directly and positively on the mRNA nodes in the network. To summarize, the interaction model contains:

1. gene nodes: the sign stems from GCN variation between two patients,
2. mRNA nodes: the sign stems from EL variation between two patients,
3. proteins: the sign stems from protein activity variation between two patients.

GCN variations and EL variations come from CGH and microarray data. The protein activity variations remain unknown but can be predicted thanks to our formalism.

3.3 Encoding variations

For each gene \( k \), we define \( GCN_{i,j}^k = CGH(i,k) - CGH(j,k) \), where \( CGH(i,k) \) is the CGH level of the gene \( k \) in the patient \( i \) smoothed by GLAD algorithm. When \( |GCN_{i,j}^k| > 0.2 \) the variation is considered as significant [6].

Similarly, for gene expression variation \( EL_{i,j}^k = EL(i,k) - EL(j,k) \), where \( EL(i,k) \) is the mean expression level measured by Affymetrix probes corresponding to the gene \( k \) in the patient \( i \). To evaluate the significance of the variation, a Student test was used on the set of probesets measuring \( EL(i,k) \) with an alpha risk of 5%.

Both for gene and mRNA nodes, significant variations are encoded + or −. The ? sign is used for nodes that are undetermined at various steps of our calculations.

3.4 Consistency analysis

For each pair of patients, we solve the system of qualitative equations (1), augmented by the information on signs coming from data. If there are solutions, the system is declared compatible. In case of compatibility some nodes have the same unique sign in each one of the many possible solutions. The unique signs of these nodes (called hard components) are predictions of the model. By this, the signs on protein nodes are predicted.

If no solution can be found, a localization of the source of conflict is attempted by subsystem analysis. First, all local violations (meaning that at least one equation (1) is violated by data information) are declared “local inconsistencies”. All locally inconsistent patterns have the same structure: one node together with its predecessors. All the other situations are declared “global inconsistencies”. Globally inconsistent patterns are more complex (they contain at least two nodes with their respective predecessors).

Notice that testing the consistency and looking for sources of conflicts is actually a NP-hard question. It appears that the topology of the network allows to
handle these questions. We used decision diagrams, a data structure meant to repre-
sent functions on finite domains; it is widely used for the verification of circuits or
network protocols. Using such a compact representation of the set of solutions, we
proposed efficient algorithms for computing solutions of the systems, predictions,
and other properties of a qualitative system [17].

3.5 Monte Carlo estimates for statistical significance of consistency

Consistency could occur also by chance. In order to estimate the significance of
consistency results, we used random perturbations and Monte Carlo estimates of
the mean numbers of pairs of patients for which random data is consistent with the
network.

For a pair of patients (i,j), let us note:
1. \(C_{i,j}^+\) and \(C_{i,j}^-\) the set of genes for which the gene copy numbers vary positively,
resp. negatively, between the patients i and j.
2. \(E_{i,j}^+\) and \(E_{i,j}^-\) the set of genes for which the gene expressions vary positively, resp.
negatively, between the patients i and j.

Straightforwardly, \(C_{i,j}^+ \cap C_{i,j}^- = \emptyset\) and \(E_{i,j}^+ \cap E_{i,j}^- = \emptyset\)

The qualitative equations (1) were solved with \(N = 1000\) data sets (each data
set contains \(P(P-1)/2\) patient comparisons, where \(P\) is the number of patients)
produced by randomly permuting the elements contained in \(C_{i,j}^+, C_{i,j}^-, E_{i,j}^+\) and \(E_{i,j}^-\). For each random dataset, consistency was tested. In case of consistency, predic-
tions on network nodes were computed. Each random dataset \(r\) is consistent \(N_r^C\)
times, locally inconsistent \(N_r^{LJ}\) times and globally inconsistent \(N_r^{GI}\) times. Note that
\(N_r^C + N_r^{LJ} + N_r^{GI} = P(P-1)/2\).

The distributions of \(N^C\), \(N^{LJ}\) and \(N^{GI}\) provide the estimates for the number of
consistent and inconsistent pairs with random data we are looking for.

4 Results

In this section, we apply our method to Ewing sarcoma data. We show results for a
couple of questions - the first concerning the relation between \(GCN_v\) and \(EL_v\); the
second concerning the model consistency tests and the impact of \(GCN_v\) on them.
4.1 Discovering links between gene copy number variations and expression level variations

How are Gene Copy Number variations (GCNv) and Expression Level variations (ELv) related?

GCNv and ELv were evaluated for each gene and each patient pair. There are 39 patients, thus 741 patient pairs.

First, the Figure 1 and Figure 2 show the repartition of the variation numbers in the patient pairs.

It is striking to see the difference in variation repartitions. GCNv are less frequent but also mainly distributed along the x and y axis. This is coherent with the relative genome stability of ES.

In spite of this general trend, some ES can exhibit a high number of GCNAs. When these unstable tumors are compared to rather stable tumors, imbalances are favored in one sense rather than the other, giving this picture with most of the pairs around the 0,0 point and distributed along the x and y axis.

A different picture can be observed with expression data. Variations are more frequent and distributed in a larger area, showing a rather homogeneous variability of ELv among the patient pairs.

From these figures, one can imagine that GCNv and ELv are independent variables.
Distribution of gene expression variations in pairwise comparisons

Fig. 2 Repartition of the ELv⁺ and ELv⁻ cardinals in the 741 patient pairs. Each point represents at least one patient pair - the number of patient pair is color-coded according to the palette on the right. The distribution is much different than the vCGH one.

To verify this, a Pearson $\chi^2$ independence test was performed under the null hypothesis that GCNv and ELv are two independent variables. The repartitions and contributions to the $\chi^2$ score are shown in Table 1.

<table>
<thead>
<tr>
<th>GCNv</th>
<th>0</th>
<th>+</th>
<th>-</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>53117</td>
<td>2197</td>
<td>2598</td>
<td>14447</td>
</tr>
<tr>
<td></td>
<td>6969</td>
<td>294</td>
<td>573</td>
<td>1948</td>
</tr>
<tr>
<td></td>
<td>6781</td>
<td>550</td>
<td>283</td>
<td>2127</td>
</tr>
<tr>
<td></td>
<td>1194</td>
<td>38</td>
<td>32</td>
<td>218</td>
</tr>
</tbody>
</table>

| Expected | 52547 | 2386 | 2701 | 14523 |
|         | 7132  | 323  | 365  | 1964  |
|         | 7101  | 321  | 364  | 1955  |
|         | 1080  | 49   | 55   | 297   |

| Contribution to $\chi^2$ | 2.59 | 15.0 | 3.98 | 0.40 |
|                         | 3.74 | 2.54 | 118.09 | 0.13 |
|                         | 14.4 | 162.91 | 17.9 | 15.10 |
|                         | 11.96 | 2.42 | 9.84 | 21.23 |

Table 1 Pearson $\chi^2$ Independence test. $\chi^2 = 402, 2 \gg 27.88$, the $\chi^2$ value for 9 freedom degrees with an alpha risk of 0.001. The major contribution is given by situations where a gene changes in an anti-correlated way in GCN and in EL. The "?" sign corresponds to cases when a probe signal in at least one experiment is too noisy to assess the variation sign.
The $\chi^2$ statistics equals 402, much greater than the value for an alpha risk of 0.1%, with 9 freedom degrees. The null hypothesis can be confidently rejected. This confirms that GCNAs can affect transcription in ways that can be investigated by comparing pairs of patients.

Moreover, the major contribution occurs when $GCN_v$ and $EL_v$ have opposite signs. This is even more striking on the whole gene set ($\chi^2 = 88761$ ; contribution of $GCN_v$ and $EL_v$ having opposite sign = 83,7%). This was clearly unexpected and it is highly counter-intuitive.

We will show that our qualitative reasoning method allows to find explanations to this surprising phenomenon.

### 4.2 Checking the EWS-FLI1 regulation model

To address this issue, we used a systems biology approach based on the qualitative analysis to confront an interaction model with CGH and expression data.

The consistency analysis raw results are shown on Figure 3.

We were concerned with two main questions:

- In which proportion is the EWS-FLI1 network model consistent with real data?
  - In cases of inconsistency, what does this tell about the model?
- Is information contained in CGH data useful to uncover regulations?

#### 4.2.1 Explaining inconsistencies

On real data including $GCN_v$ influences, the model is consistent with the data in 317 patient pairs (42.8% of the 741). Additionally, 314 (42.4%) local inconsistencies and 110 (14.8%) global inconsistencies were found.

Understanding the incompatibility sources may help to focus on the model weaknesses. First, it is necessary to analyze the sources of local inconsistencies, the most numerous ones.

All the local inconsistencies have the same origin : a patient pair $(i,j)$ where there is at least one gene $k$ for which : $GCN_v(i,j,k) = -EL_v(i,j,k)$. We call this an anticorrelated variation.

This is not a rare case: from the Table 1, there are 1123 cases in the 741 patient pairs. They are spread in 367 patient pairs and involve 67 genes (see Supplementary_Table_1).

Hence, 367 local inconsistencies were expected. This means that 50 pairs that were expected to be locally inconsistent were explained by the model.

More precisely, on the 67 genes that are involved in anticorrelated variations, 23 are never involved in local consistencies (see green colored genes in the Supplementary_Table_1). This is due to the presence in the network model of at least one transcription regulation on those genes. Those explained locally inconsistent influences appear 414 (36.9%).
In other words, local inconsistencies point to the lack of transcription regulations in the model. Adding them can potentially remove all local inconsistencies.

The global inconsistencies point to other model weaknesses. Unfortunately, our solver is not able to localize the whole set of inconsistent subgraphs (see section 2)\(^3\). Only 33 influences are involved in the inconsistent set we obtained. All of them are implied in transcription regulation, including: TP53, E2F1 and EWS – FLI1 (see Supplementary Table 2). Therefore, our method allows to focus on subgraphs of a complex model that need refinement to become consistent with observations.

4.2.2 CGH data increase the consistency between the data and the model

To assess what relevant information is contained in CGH data, it can be useful to:

- reproduce the consistency analysis without taking into account the anticorrelated variations that remain unexplained by the model;
- compare the result to the consistency analysis when GCN influences are removed (by taking into account only ELv, hence discarding CGH information).

\(^3\) Notice that a more powerful implementation of constraints solver, with Answer Set Programming, will be soon available to overcome this technical problem.
This analysis gave the results shown in Figure 3. Without CGH information, the model is consistent with $EL_v$ alone in 525 (70.9%) patient pairs. Using CGH information on the genes having transcription regulators, the model is consistent with $EL_v$ and $GCN_v$ in 528 (71.3%) patient pairs.

In 3 cases, the model becomes consistent thanks to information from CGH data. To understand better the impact of $GCN_v$ influences, the 3 inconsistent subgraphs were represented using Cytoscape software. The Figures 4 and 5 show an example of this analysis. The other cases are shown in Supplementary Figures 1 and 2.

On this example, the inconsistency is localized on the mRNA $IGFBP_3$ node. $IGFBP_3$ is positively targeted by a set of regulators, except $EWS-FLI_1$ that was shown to inhibit $IGFBP_3$. However, between the two patients, given the observations, mRNA $IGFBP_3$ should be activated. This is not the case, producing the inconsistency.

The $IGFBP_3GCN_v$ proposes an explanation to this phenomenon: due to its negative variation between the two patients, the negative $IGFBP3EL_v$ can be understood. A biological interpretation for such a pattern can be: despite the positive signals arising from various regulators, the difference in gene copy number is sufficient to decrease $IGFBP3EL$.

Obviously, this hypothesis must be confirmed by experimental validation.

To conclude, CGH data bring information on local variations that have an influence on the $EWS-FLI_1$ network model.

### 4.2.3 Assessing the statistical quality of consistency frequencies

In order to assess the quality of consistency tests, a randomization of input data was performed using 1000 random permutation on $GCN_v$ and $EL_v$ for each patient pair.
The 741000 data sets with GCNv were confronted to the EWS-FLI1 model; the same analysis was repeated on the same datasets without considering GCN influences.

Consistency analyzes with and without GCN influences were performed to be compared to results on real datasets. Results are exposed in Figure 3. This shows that there are less consistencies and proportionally more inconsistencies on random data than on real data.

The distribution of consistency frequency distribution obtained for the 1000 datasets including GCN influences follows a normal distribution ($\mu = 279$, $\sigma = 10.1$ - Kolmogorov-Smirnov normality test value = 0.0289 < 0.0386 , the bilateral value for an alpha risk of 5%).

Given such a distribution, the probability to obtain a consistency frequency equal to or greater than 317$^4$ equals 3.79%.

Similarly, the distribution of the consistency frequency distribution obtained on the 1000 datasets without GCN influences (see Figure 6). The distribution follows a normal distribution ($\mu = 446$, $\sigma = 12.5$ - Kolmogorov-Smirnov normality test value = 0.0251 < 0.0386 , the bilateral value for an alpha risk of 5%).

Given such a distribution, the probability to obtain a consistency frequency equal to or greater than 525 $^5$ equals $1.31 \times 10^{-10}$. This probability is even lower for the real data set using GCN explained by the model$^6$.

This proves that one can trust the consistency frequency obtained on real data sets.

Fig. 5 Subgraph with GCN influences on patient pair (EW57,EWS8). Observed signs of gene expression variation are circled. The IGFBP3 GCN varies negatively between the two patients and resolves what was previously an inconsistency between the model and ELv alone.

$^4$ the consistency frequency obtained on real data set without GCN.

$^5$ The consistency frequency obtained on real data set without GCN.

$^6$ The consistency frequency obtained on real data set with the GCN influences explained by the model equals 528.
Fig. 6 Consistency frequency distribution for randomized data without \(CGH\) influences. The distribution follows a normal distribution \((\mu = 446, \sigma = 12.5)\). It must be compared to the consistency number obtained on real data without \(GCN_v\) influences (525) and with \(GCN_v\) explained by the model (528).

However, it is surprising to observe such a high number of consistent cases on randomized data sets. We are currently investigating the reasons. Two hypotheses motivate us:

- the network topology may be robust to random variations;
- there is an effect of the number of constraints imposed by observations - as there is a variability in \(|GCN_v|\) and \(|EL_v|\) as shown on Figure 1 and 2.

A simple way is to test whether a significant correlation exists between \(|GCN_v(i, j)|\) and \(|GCN_v(i, j)|\) and the consistency for each patient pair \((i, j)\).

5 Discussion

Given the difficulties to analyze CGH data on ES tumors, we propose to change of paradigm and propose a systems biology approach dedicated to investigate the inter-patient variability simultaneously at genomic and transcriptomic levels and their compatibility with a EWS-FLI1 gene regulation network.

This study addresses two main issues: 1) the link between CGH and expression pairwise variations in 39 ES; 2) the consistency between a \(EWS - FLI1\) model and these variations.
To handle the first question, interesting representations of patient pairs as functions of \( GCN_v \) and of \( EL_v \) cardinalities were produced. It appears that the patient pair distribution following \( GCN_v \) is highly different from its counterpart following \( EL_v \). This shows that patient pairwise comparisons exhibit different transcriptomic and genomic variability patterns. One could be mistaken in interpreting this as the result of an independence between the variables \( GCN_v \) and \( EL_v \).

The \( \chi^2 \) independence test states that these two variables are undoubtedly related. This agrees with biological intuition: when a gene copy number increases, the gene expression level is expected to increase - and vice versa.

However, surprisingly, the major dependency contribution comes from anticorrelated variations. This is true for the whole set of measured genes. This suggests the existence of a feedback regulation of genes present in altered regions that counteracts \( GCN \) imbalances.

The \( EWS - FLI1 \) network model is able to deal in part with these anticorrelated variations: no gene having at least one transcriptional regulation appears in local inconsistencies. On the contrary, if a “deficient gene” does not have a transcription regulation, it will be involved in a local inconsistency if its \( GCN_v \) and \( EL_v \) appear anticorrelated. Thus, our method points to the model incompleteness. Adding missing transcription regulations can potentially remove all local inconsistencies.

To answer the second question, the compatibility of the \( EWS - FLI1 \) model with the pairwise genomic and transcriptomic variations was verified. It appears that the model is consistent with expression data in more than 70% of the cases after having silenced the “deficient gene” \( GCN_v \) influence.

3 cases that were inconsistent using \( EL_v \) alone become consistent. The analysis of these inconsistent subgraphs shows that some \( EL_v \) that were unexplained by the model could be explained by local \( GCN \) variations. This suggests that local \( GCN \) variations carry valuable information that can propagate through the interactions. This result validates the capacity to investigate such local effects of \( GCN_v \) by our approach.

Finally, we compared our results to 1000 randomized datasets. It appears that the consistency frequencies obtained on real datasets cannot be obtained by chance.

Another intriguing phenomenon appeared during this latter analysis: we did not expect so high consistency frequencies on randomized datasets. We are currently studying whether this is related to the genomic and transcriptomic variation number or whether this is a consequence of the intrinsic network robustness.

\textit{Biological system robustness} may be the key to understand apparent contradictions in experimental data on ES. Let us consider the following paradox: the existence of a general trend that relates genetic instabilities to worst prognosis is opposed to the difficulty of finding repeated and specific genetic disorders linked to tumor outcomes.

If we consider that genetic instability acquisition is a stochastic process, stemming from a disturbed DNA repair machinery, it is likely that the largest part of genetic disorders have no individual effect on the cell physiology. This may result from a negative feedback control.
In the same time, it is also possible that, in exceptional cases, a specific gene disorder manages to overcome feedback and have visible effects on cell physiology. EWS-FLI1 itself is an extreme example.

As future work, we intend to use our method to detect these exceptional cases. We already proved that in a very limited number of cases (3 on 216 inconsistencies) the information on GCNv carried by specific genes can explain an unusual network behavior.

The novel hypothesis that outcomes from this work is that genetic disorder accumulation can have a global impact by increasing the probability that a specific gene disorder has consequences on a stabilized network. We expect that such events will be found more frequently in metastatic tumors than in non metastatic ones.

6 Acknowledgements

We acknowledge the financial support from ANR BIOSYS-2006 program (SITCON project) and the PIC Bioinformatique et Biostatistiques from Institut Curie. AZ, EB and GS are members of the team “Systems Biology of Cancer”, équipe labellisée par la Ligue Nationale Contre le Cancer. We thank Philippe Hupé and Philippe La Rosa for help and disponibility for using Vamp. We also thank Michel Le Borgne and Jacques Nicolas for stimulating discussions and the bioinformatics GenOuest platform for their invaluable help to overcome technical problems and use the Bioquali application efficiently. SB’s work was supported by a INRIA-French ministry postdoc grant.

References


Revealing Disease Mechanisms via Coupling Molecular Pathways Scaffolds and Microarrays: A Study on the Wilm’s Tumor Disease

Alexandros Kanterakis\textsuperscript{1,3}, Vassilis Moustakis\textsuperscript{1,3}, Dimitris Kafetzopoulos\textsuperscript{2}, and George Potamias\textsuperscript{1}

\textsuperscript{1}Institute of Computer Science, FORTH, Greece
\textsuperscript{2}Institute of Molecular Biology and Biotechnology, FORTH, Greece
\textsuperscript{3}Department of Production Engineering and Management, TUC, Chania, Greece

Abstract. Moving towards the realization of genomic data in clinical practice, and following an individualized healthcare approach, the function and regulation of genes has to be deciphered and manifested. This is even more possible after the later advances in the area of molecular biology and biotechnology that have brought vast amount of invaluable data to the disposal of researchers. Two of the most significant forms of data come form microarray gene expression sources, and gene interactions sources – as encoded in Gene Regulatory Pathways (GRPs). The usual computational task involving microarray experiments is the gene selection procedure while, GRPs are used mainly for data annotation. In this study we present a novel perception of these resources. Initially we locate all functional paths encoded in GRPs and we try to assess which of them are compatible with the gene-expression values of samples that belong to different clinical categories (diseases and phenotypes). Then we apply usual feature selection techniques to identify the paths that discriminate between the different clinical phenotypes providing a paradigm shift over the usual gene selection approaches. The differential ability of the selected paths is evaluated and their biological relevance is assessed. The whole approach was applied on the Wilm’s tumor domain with very good and indicative results.

1. Introduction

The interdisciplinary research field of molecular biology and bioinformatics is continuously enriched by the advances in many areas such as sequence analysis, genome annotation and analysis of gene and protein regulation. These advances have brought to the present the post-genomic era where, as the basic knowledge is tamed, we are mainly seek for methods that integrate various and heterogeneous
types of established biological knowledge. The major question to deal with relates to the regulation of the function of genes, targeting the ways that this function affects the overall phenotype of a living organism. A first step towards this is the combined processing of clinical and genomic data. Clinical data as the explicit expression of the phenotypic features of an organism should be integrated with the genomic data that represent the genotypic signature of the organism. This effort can help researchers to gain insights about the role of gene function in pathology, locate the risks and susceptibilities of each unique person and thus, provide individualized healthcare [1].

From a biology point of view, the goal is to provide a systematic, genome-scale view of genes interactions and functionality [2]. The advantage of this approach is that it can identify emergent properties of the underlying molecular system as a ‘whole’ – an endeavor of limited success if targeted genes, reactions or even molecular pathways are studied in isolation [3]. Individuals show different phenotypes for the same disease – they respond differently to drugs and sometimes the effects are unpredictable. Many of the genes examined in early clinico-genomic studies were linked to single-gene traits, but further advances engage the elucidation of multi-gene determinants of drug response. Differences in the individuals’ background DNA code but mainly, differences in the underlying gene regulation mechanisms alter the expression or function of proteins being targeted by drugs, contribute significantly to variation in the responses of individuals. The challenge is to accelerate our understanding of the molecular mechanisms of these variations and to produce targeted individualized therapies.

In this paper we present an integrated methodology that couples and ‘amalgamates’ knowledge and data from both Gene Regulatory Pathway (GRP) and Microarray (MA) gene-expression sources. The methodology comprises two main parts. In the first part we decompose a number of targeted pathways – pathways involved in particular disease phenotype, into all possible functioning paths (i.e., part of a molecular pathway). Then, by introducing gene expression knowledge from a MA experiment we rank all paths according to the ‘compatibility difference’ they exhibit among the samples of different clinical phenotypes. In the second part of the methodology we substitute genes with paths and gene expression with compatibility value ranks. At the end we apply feature selection techniques to identify those functional paths that differentiate between the targeted phenotypic classes, and we assess their prediction power (classification) performance. As a proof of concept we apply the technique on a microarray experiment that targets the Wilms’ tumor (WT; nephroblastoma) disease. We were able to identify significant paths in various molecular pathways that reveal distinct mechanisms between different WT phenotypes. The targeted WT phenotypes concern the tumor grade histological feature. Results are discussed about their biological relevance.

A preliminary implementation of the methodology is made in a system called MinePath. MinePath aims to uncover potential gene-regulatory ‘fingerprints’ and mechanisms that govern the molecular and regulatory profiles of diseases.
2. MAs and GRNs as sources of biomedical knowledge

2.1. Microarrays

Microarrays [4], [5] are devices able to measure simultaneously the expression of thousands of genes, revolutionizing the areas of molecular diagnostics and prognostics. A number of pioneering studies have been conducted that profile the expression-level of genes for various types of cancers such as leukemia, breast cancer and other tumors [6], [7]. The aim is to add molecular characteristics to the classification of diseases so that diagnostic procedures are enhanced and prognostic predictions are improved. These studies demonstrate the great potential and power of gene-expression profiling in the identification and prediction of various disease phenotypes and prognostic disease factors.

Gene-expression data analysis depends on Gene Expression Data Mining (GEDM) technology, and the involved data analysis is based on two basic approaches: (a) hypothesis testing - to investigate the induction or perturbation of a biological process that leads to predicted results, and (b) knowledge discovery - to detect underlying hidden-regularities in biological data. For the latter, one of the major challenges is gene-selection [8], [9]. Possible prognostic genes for disease outcome, including response to treatment and disease recurrence, are then selected to compose the molecular signature (gene-markers) of the targeted disease.

2.2. Gene Regulatory Pathways

GRPs are network structures that depict the interaction of DNA segments during the transcription of genes into mRNA. The prominent and vital role of GRPs in the study of various biology processes is a major sector in contemporary biology research, where numerous thorough studies have been conducted and reported [10], [11]. From a computational point of view, GRPs can be conceived as analogue of biochemical computers that regulate the level of expression of target genes [12]. Each network has inputs, usually proteins or transcription factors that initiate the network function. The outputs are usually certain proteins (encoded by specific genes). The network by itself acts as a mechanism that determines cellular behavior where the nodes are genes and edges are functions that represent the molecular reactions between the nodes. These functions can be perceived as Boolean functions, where nodes have only two possible states (“on” and “off”), and the whole network being represented as a simple directed graph [13]. The notion of GRPs is by itself an abstraction of the underlying chemical dynamics of the cell, thus the expectation of high reliability in terms of modeling is limited. It is indicative that most of the relations in known and established GRPs have been derived from laborious and extensive laboratory experiments and careful study of the existing biochemical literature. Thus GRPs are far from being complete, at least with
respect to their ability to capture and model all the internal cell dynamics of complex living organisms.

Current efforts focus on the reconstruction of GRPs by exploring gene-expression data. For example in [14] it is reported that network topologies, as extracted from gene co-expression events, could discover motifs and regulatory hubs that can characterize the entire cellular states and guide further pharmaceutical research. Very few methods of gene regulatory inference are considered superior, mainly because of the intrinsically noisy property of the data, ‘the curse of dimensionality’, and the lack of knowledge about the ‘true’ underlying structure of the networks.

The study of the function, structure and evolution of GRPs in combination with microarray gene-expression profiles and data is essential for contemporary biology research. First of all, researchers have uncovered a multitude of biological facts, such as protein properties and genome sequences. But this alone is not sufficient to interpret biological systems and understand their robustness, which is one of the fundamental properties of living systems at different levels [15]. This is mainly because cell, tissues, organs, organisms or any other biological systems defined by evolution are essentially complex physicochemical systems. They consist of numerous dynamic networks of biochemical reactions and signaling interactions between active cellular components. This cellular complexity has made it difficult to build a complete understanding of cellular machinery to achieve a specific purpose [16]. To circumvent this complexity microarrays and molecular networks can be combined in order to document and support the detected and predicted interactions [17]. The advances and tools that each discipline carries can be integrated in a holistic and generic perspective so that the chaotic complexity of biology networks can be ‘screened’ and traced down.

2.3. Coupling MAs and GRPs

Microarray experiments involve more variables (genes) than samples (patients). This fact, leads to results with poor biological significance. There is an open debate whether we should concentrate on gathering more data or on building new algorithms in order to improve biological significance. Simon et al. in [18] published a very strict criticism on common pitfalls on microarray data mining while in [19] comments about the bias in the gene selection procedure are presented. Moreover, due to limitations in DNA microarray technology higher differential expressions of a gene do not necessarily reflect a greater likelihood of the gene being related to a disease and therefore, focusing only on the candidate genes with the highest differential expressions might not be the optimal procedure [20]. Another significant aspect is the noisy content microarray experiments. Appropriate statistical analysis of noisy data is very important in order to obtain meaningful biological information [21], [22]. Evidence on this is given by the fact that different methods produce gene-marker lists that are strikingly different [23]. As a re-
sult, and because the immature state of microarray technology, reproducibility of microarray experiments and the accompanied statistical prediction models are pretty low, except when protocols are uniformly and strictly followed [24], [25].

In the light of the aforementioned observations and in order to overcome the posted limitations we have to consider MA-based gene-expression profiles just as an instance of biological information, strongly connected - rather than isolated, from other sources of related biological knowledge. In other words, gene-expression profiles should be examined, explored and interpreted not as ‘static’ but as instances of the underlying regulatory framework, as encoded by established and known GRPs.

3. Methodology

Existing GRPs databases provide us with widely utilized networks of proved molecular validity. The most known are network that describe important cellular processes such as cell-cycle, apoptosis, signaling, and regulation of important growth factors. Online public repositories contain a variety of information that includes not only the network per se but links and rich annotations for the respective nodes (genes) and edge (regulation). In the current study we utilize the KEGG pathways repository*. KEGG provides a format representation standardized by its own markup description language (KGML†).

The gene regulatory relations we consider are restricted to what might be observed in a microarray experiment: a change in the expression of a regulator gene modulates the expression of a target gene mainly via protein-DNA interactions. In other words, there are genes that causally regulate other genes. A change in the expression of these genes might change dramatically the behavior of the whole network. The identification and prediction of such changes is a challenging task in bioinformatics. Moreover, we have to identify real, true networks and use them as scaffolds [26] to methods that infer gene regulatory networks out from gene expression data. This approach can aim several areas of biology research such as genomic medicine [27], microarray data mining [28] and phylogenetic analysis [29]. We have implemented our approach on coupling GRPs and MA data in a system called MinePath.

3.1. Pathway decomposition

MinePath relies on a novel approach for GRP processing that takes into account all possible functional interactions of the network, the network’s scaffolds. The different GRP scaffolds correspond to the different functional paths that can be followed during the regulation of a target gene.

* KEGG: Kyoto Encyclopedia of Genes and Genomes; http://www.genome.jp/kegg/
† KGML (KEGG Markup Language); http://www.genome.jp/kegg/xml/
Fig. 3.1. Function-path decomposition – the GRP scaffolds: Top: A target part of the KEGG cell-cycle GRP; Bottom: The five decomposed fictional paths (scaffolds) for the targeted path part – all possible functional routes taking place during network regulation machinery.

Different GRPs are downloaded from the KEGG repository. With an XML parser (based on the specifications of the KGML representation of GRPs) we obtain all the internal network semantics (see next sub-section). In a subsequent step, all possible and functional network paths are extracted as exemplified in Fig 3.1. Each functional path is annotated with the possible valid values according to Kauffman’s principles that follow a binary setting: each gene in a functional path can be either ‘ON’ or ‘OFF’. According to Kauffman [13], the following functional gene regulatory semantics apply: (a) the network is a directed graph with genes (inputs and outputs) being the graph nodes and the edges between them representing the causal (regulatory) links between them; (b) each node can be in one of the two states: ‘ON’ or ‘OFF’: ‘ON’ and ‘OFF’ states correspond to the gene being expressed (i.e., the respective substance being present) or not expressed, respectively; and (c) time is viewed as proceeding in discrete steps - at each step the new state of a node is a Boolean function of the prior states of the nodes with arrows pointing towards it.

Since the regulation-edge connecting two genes defines explicitly the possible values of each gene, we can set all possible state-values that a gene may take in a path. Thus, each extracted path contains not only the relevant sub-graph but the state-values of the involved genes as well. The only requirement concerns the following assumption: for a path being functional it should be ‘active’ during the GRP regulation process; in other words we assume that all genes in a path are functionally active. For example, assume the functional path A → B (‘→’ is an activation/expression regulatory relation). If gene A is on an ‘OFF’ state then, gene B is not allowed to be on an ‘ON’ state - B could become ‘ON’ only and only if it is activated/expressed by another gene in a different functional path (e.g., C → B). The assumption follows a ‘closed world assumption’, that is: if what we know is
just the ‘A $\rightarrow$ B’ gene-gene interaction then, B could be activated only from A; if A is inactive there is no causal evidence for B being active. If we had allowed non-functional genes to have arbitrary values then the significant paths would be more likely to be ‘noisy’ rather than exhibiting some form of biological importance.

After parsing the targeted GRPs, the involved genes are stored in a database that acts as a repository for future reference. Through this repository we can query paths being parts of target GRPs, GRPs that contain specific genes or target a specific regulatory relation. Moreover, the stored paths can be combined and analyzed in the view of specific microarray experiments and respective gene-expression sample profiles. As the database repository contain and retrieves functional paths from a variety of different GRPs (e.g., cell-cycle, apoptosis etc), we may combine different molecular pathways and networks – a major need for molecular biology and a big challenge for systems biology and contemporary bioinformatics research.

### 3.2. Combining gene-expression profiles and functional paths

The next step is to locate microarray experiments and respective gene-expression data for which we expect (suspect) the targeted GRPs play an important role. For example the cell-cycle and apoptosis GRPs play an important role in tumorgenesis and cancer progression.

With a gene-expression/functional-path matching operation, the valid and most prominent GRP functional paths are identified. These paths uncover and present potential underlying gene regulatory mechanisms that govern the gene-expression profile of the samples under investigation. Such a discovery may guide to the finer classification of samples as well as to the re-classification of diseases, providing the most prominent molecular evident for that.

### 3.3. Matching GRP paths with MA data

The samples of a binary transformed (discretized) gene-expression matrix are matched against targeted molecular pathways and respective GRP functional paths (retrieved form the described repository). We follow a gene-expression value discretization process presented elsewhere (please refer to [9]). As already exemplified, GRP and MA gene-expression data matching aims to differentiate GRP paths and identify the most prominent functional paths for the given samples. In other words, the quest is for the paths that exhibit high matching scores for one of phenotypic class and low matching scores for another. This is a paradigm shift from mining for genes with differential expression to mining for subparts of GRP with differential function. The algorithm for differential path identification is inherently simple (see Fig. 3.3).
Fig. 3.3. Matching Functional-paths (scaffolds) and gene-expression profiles: Samples S1, S2, S3 belong to the ‘+’ class and samples S4, S5 belong to the ‘−’ class. The first path (IL-1R → TRADD) satisfies samples 1,2,3,5. Second path (IL-1R → TRADD → FLIP) satisfies samples S1, S2, S3. Third path satisfies all samples and the forth path doesn’t satisfy any sample. The green arrow indicates that the second path yields the maximum differential power and it contains a potential function differentiation since it is consisted only with samples that belong to the ‘+’ class. (‘’; activation; ‘’; inhibition).

For each path we compute the number of samples that is consistent for each disease phenotypic class. Suppose that there are $S_1$ and $S_2$ samples belong to the two classes, respectively. Assume that path $P_i$ is consistent with $S_{i;1}$ and $S_{i;2}$ samples form the first and second class, respectively. Formula 1,\

$$\left| \left( S_{i;1} / S_1 \right) - \left( S_{i;2} / S_2 \right) \right|$$

computes the differential power of the specific path with respect to the two classes. Ranking of paths according to formula 1 provides the most differentiating and prominent GRP functional paths for the respective disease phenotypes. These paths present evidential molecular mechanisms that govern the disease itself, its type, its state or other targeted disease phenotypes (e.g., positive or negative response to specific drug treatment). The formula can be enriched so that longer consistent paths acquire stronger power. It can also be relaxed so that ‘consistent’ is a continuous indicator rather than a Boolean value. Finally we may introduce ‘unknown’ values for missing and erroneous gene expression values.

4. Revealing Regulating Wilm’s tumor Molecular Mechanisms

The presented MA-GRP coupling methodology was applied on a study for expression profiling of the Wilm’s tumor (WT, nephroblastoma) disease [30]. In the original publication the researchers report new candidate genes for various WT clinical phenotypes.
WT samples were divided according to the histological risk grade ('low/intermediate' and 'high'), relapse of tumor ('no', 'yes'), survival ('relapse-free' and 'death'), metastasis ('no', 'yes') and response to chemotherapy ('good', 'poor'). The results presented in this paper focus on the histological risk grade as a target WT phenotype. In the original published study a set of 20 differentially expressed genes are reported for this WT phenotype [30].

From the ArrayExpress online microarray experiments' repository (http://www.ebi.ac.uk/microarray-ae/) we downloaded the expression values and the clinical annotation of 138 samples from the WT study - 108 of them being classified as histology risk 'low/intermediate', and 30 as 'high'. For this study we targeted 17 GRPs – the selection was made on the basis of their susceptibility and incrimination to the WT disease and on established biological and clinical knowledge of their involvement in cell regulatory tumor growth mechanisms. The path decomposition process resulted into a total of 8937 functional paths. Most of these paths didn’t show any special differential ability over the samples. In order to identify the significant paths the matching gene-expression formula (formula 1 presented in section 3.3) was applied. A threshold value of 0.5 was set to filter-out not differential paths (the threshold was fixed after experimentation with various cut-off values). Filtering resulted into a set of 87 functional-paths for further exploration.

The next step was to find the most relevant and discriminant functional-paths, and build a classifier able to distinguish between the two phenotypic classes - 'high' and 'low' (including 'intermediate' samples) histological risk grade, respectively. The whole dataset is presented as a binary-[0,1] array-matrix of 87 lines for functional paths, and 138 columns for samples. The value ‘1’ in the position $i,j$ of this array means that the $i$ path is ‘active’ for sample $j$. Active means that all genes that comprise this path are either ‘ON’ or ‘OFF’ according the interaction relationships of the genes of the path. Respectively, a ‘0’ value means that the genes involved in the path do not exhibit the same value as the expression value of the respective sample. The array-matrix can be seen as an indicator of which paths are functional on which samples. Furthermore, it comprises a resemblance of normal gene-expression matrices - instead of genes being either active or inactive, according to their expression over different kind of samples, we have paths being functional or non-functional over the same set of samples. This gives us the ability to apply whatever feature selection processes to select the most relevant and discriminant functional-paths. For this, we rely on a feature/gene-selection algorithmic process presented in [9].

Initially a Wilcoxon rank-sum test ($p<0.005$) was applied that reduced the functional-paths from 87 to 54. Then, ranking and selection of the most discriminant functional-paths was performed – ranking is based on an information-theoretic entropic formula, and selection encompasses a naïve Bayes classification process [9]. The whole process resulted into a complex of four discriminant and indicative functional-paths (see Fig. 4.)
The four selected functional-paths are involved in two different GRPs: one from ‘VEGF signaling’, and three from ‘Focal adhesion’. The three functional-paths from ‘Focal adhesion’ are subparts of the whole ‘Focal adhesion’ GRP.

We performed a Leave One Cross Validation (LOOCV) procedure in order to assess the discrimination/classification performance that these paths exhibit (note that each functional-path is now considered as a feature). A 95% LOOCV (131/138 samples) accuracy figure was achieved when, the fitness (i.e., train vs. train) figure was inferior, 91% (126/138, 8 misses for ‘high’ classified as ‘low’, and 4 misses for the inverse case). This finding is quite interesting: beside the high accuracy performance data ‘overfitting’ is reduced. This is a strong indication for the high relevance of the four identified functional-paths (at least for the available dataset).

In addition, we applied the same feature/gene selection algorithmic process on the original gene-sample matrix (i.e., the normal gene-selection setting for microarray gene-expression profiles). This resulted into the same LOOCV accuracy and in 89% (123/128) fitness (17 genes selected) accuracy figures. A potential speculation on this finding is the following: with the presence of thousands of genes and of a limited number of samples, gene-selection processes are lean to overfitting events. We believe that this explains the diversity of results produced by available gene-selection techniques, their instability on different population (for the same disease) cohorts, and the inability to relate statistical significance with biological relevance. In contrast, the introduced methodology is able to identify not just discriminant gene-markers but, discriminant, indicative and ‘stable’ gene-regulatory mechanisms that govern disease phenotypes and clinical manifestations.

In a preliminary attempt to find biological evidence for our findings we focus on the involvement of ‘P13K’ and ‘AkT’ gene-products in both identified GRPs (see Fig. 4). The related literature report the ‘P13K/AkT’ complex to be implicated in WT disease, as well as it is the main component of WT therapeutic targets [31]. Certainly, further biological validation of the approach is needed, a task for future research.
Conclusions

We have presented an integrated methodology for the coupling of both GRPs and MA gene expression profiles. In the heart of the methodology is the decomposition of GRPs into functional-paths (or, scaffolds), the matching of these paths with samples’ gene expression profiles, and the application of feature selection techniques for the identification of the most relevant and discriminant ones.

Application of the methodology on gene-expression data for the Wilms’ tumor disease showed that: we can identify a limited number of functional-paths that exhibit significantly differential behavior between different WT phenotypes (‘low/intermediate’ vs. ‘high’ histological grade risk). The findings provide valuable insights for further research over the function and role of the involved genes and their underlying regulatory machinery.

Among others, our on-going and future R&D work include: (a) further experimentation with various real-world microarray studies and different GRP targets (accompanied with the evaluation of results form molecular biology and clinical research experts); (b) extension of path decomposition to multiple GRPs; (c) elaboration on more sophisticated path/gene-expression profile matching formulas and operations; (d) incorporation of different gene nomenclatures in order to cope with microarray experiments from different platforms and nomenclatures; and (e) porting of the whole methodology in a Web-Services and scientific workflow environment.

Acknowledgements: This work was supported by the European Commission’s Sixth Framework Programme in the context of the ACGT (FP6-2005-IP-026996) Integrated project.

References


A Simple Algorithm Implementation for Pattern-Matching with Bounded Gaps in Genomic and Proteomic Sequences, on the Grid EGEE Platform, using an intuitive User Interface

Vegoudakis K. I., Margaritis K. G., Maglaveras N.

Abstract In the last decade an unprecedented development in bioinformatics has been observed. An extremely high number of organisms have been sequenced and included in genomic databases. The huge amount of data produced needs to be stored and processed for further analysis. Scientists, have researched algorithms for finding complicated patterns in DNA sequences, but there is a need for computational power and large storage systems, in order to implement specific algorithms in as many as possible DNA sequences stored in large Databases and save the results produced for future use.

Recently, a number of pattern matching algorithms that allowing gaps have been introduced and Grid is an emerging technology that seems to be helpful in this kind of biomedical research. In this paper, we present our effort towards the construction of a user friendly Interface for accessing the Grid EGEE platform. The Interface is a specific one and was built to perform Pattern-Matching with Bounded Gaps in Genomic and Proteomic Sequences. The algorithm and the Interface is tested with small and large DNA and protein sequences as an input, downloaded from a gene and protein repository, NCBI (National Center for Biotechnology Information) and the gathered results produced are presented.

Vegoudakis Konstantinos
Parallel and Distributed Processing Laboratory,
Department of Applied Informatics, University of Macedonia, 156 Egnatia str., P.O. Box 1591, 54006 Thessaloniki, Greece, e-mail: kostasve@uom.gr

Margaritis Konstantinos
Parallel and Distributed Processing Laboratory,
Department of Applied Informatics, University of Macedonia, 156 Egnatia str., P.O. Box 1591, 54006 Thessaloniki, Greece, e-mail: kmarg@uom.gr

Maglaveras Nikos
Lab of Medical Informatics, Health Sciences Faculty, Aristotle University, AUTH PB 323, 54124 Thessaloniki, Greece, e-mail: nicmag@med.auth.gr
1 Introduction

Bioinformatics is playing an increasingly large role in the study of fundamental biomedical problems and represents a new, growing area of science that uses computational approaches to answer biological questions [1]. The explosion of sequence and structural information available to researchers need to be processed and stored in special supercomputing infrastructures and storage systems respectively. Grid is an emerging technology for distributed computing in advanced science and engineering and was firstly developed as a concept to enable resource sharing within scientific collaborations [2]. Although, great investments have been made the previous years worldwide towards the construction of Grid Infrastructures, the Grid community expects to evaluate the necessity and the current potential of using the Grid [3].

Grid technologies have enormous potential for heavy computational and storage demanding applications but a few of them have been implemented and executed in the context of Grid computing. There are a lot of reasons of the slow take up of Grid computing but the most substantial one is the lack of an existing simple unique framework and easy to use User Interface for executing applications either biomedical or not. For example the gLite WMS (Workload Management System) demands the experience of using Unix-like scripts in order to interact with the Grid infrastructure.

The potential of Grid computing in healthcare has been examined and worked out by the HealthGrid initiative [4] according to which the prospects of Grid Computing are deployed e.g. computational models of systems/organisms, pharmaceutical research, large-scale epidemiological studies and so on. The scale and the complexity of the European EGEE (Enabling Grid for E-sciencE) project and the launch and the expanding of biomedical applications in it, is described in [5]. It is worth mentioning the BIOMED Virtual Organization (VO) which is hosted on the EGEE project and created in the context of the BioinfoGRID project, whose purpose is to promote the Bioinformatics applications for life science, in order to carry out research based on the Grid networking technology [6].

User-friendly access to the grid environment is of great importance for the scientific community. In particular a lot of effort has been made to construct an interface which is easy to use. The g-Eclipse project aims to build an integrated workbench framework to access the power of existing Grid infrastructures. g-Eclipse tries to ease the execution of the existing applications that are needed to be executed in the Grid environment and it provides tools for the customization of the Grid users’ applications and management of the Grid resources [7]. GuiGen is a comprehensive set of tools for creating customized graphical user interfaces and was originally designed for the use in computational grids [8]. In addition, P-GRADE provides a high-level graphical environment to develop parallel applications transparently both for parallel systems and the Grid and it also supports workflow definition and coordinated multi-job execution for the Grid. One of the main advantages of P-GRADE is that the user does not have to learn the different APIs for parallel systems and the Grid. The current version of P-GRADE supports the interactive execution of parallel
programs both on Globus-2 and Condor Grids [9]. However, the above mentioned approaches still require technical knowledge from the end users.

In this work, which is a sequence of previous work on the research of an intuitive user-friendly and generic User Interface [10], [11], either with the help of XML or not, another user-friendly interface for Pattern-Matching with Bounded Gaps in genomic sequences is presented. This interface is a specific one and it is used only for submission of certain type of jobs, those for string pattern matching with bounded gaps. The user has the ability to interact with the UI for the job submission, job status retrieval, upload/download of DNA and protein sequence files that are needed, etc.

The problem of fast searching of patterns that contain Classes of characters and Bounded size Gaps (CBG) in text occurs in various fields and the most important one, is protein matching. The design of two new practical CBG practical algorithms that are faster and simpler than all regular expression search techniques are described in [12]. In Crochemore et. al. [13], algorithms for several versions of approximate string pattern matching with gaps are presented. Further restrictions to the gaps are introduced in [14] with lower and upper bounds restrictions on the gaps. The user interface created, implements these algorithms on the Grid platform.

The applicability of the user interface is examined via a set of jobs. Different DNA and Protein sequences were used to search in them for different patterns with bounded gaps. It has to be noted that especially large DNA sequences need a lot of computing time and storage in order to execute. For this reason the interface built, enables the user to submit multiple executions of jobs with string pattern matching with bounded gaps. This method is known as parameter study and it is often met in biomedical applications.

More precisely, the user interface is expected to utilize existing bioinformatics applications on available grid testbed (such as EGEE, NGS, etc), [15]. In its final form it will require only Java. As a result installation and configuration of specific operating system and grid middleware toolkit will not be necessary. The user interface current production status, simplifies the job submission process on EGEE grid infrastructure, which is not a trivial task from a biologist’s end-user point of view. The parameter study as mentioned above, gives the biologist an assistance in his work, as he gains time, retrieving all the results from the experiments in a reasonable amount of time.

This paper is structured as follows. First, the methodology adopted is presented, providing the necessary implementation and introductory details about the string pattern matching problem with bounded gaps and the creation of the UI for the Grid. Then, the results obtained from the execution of multiple jobs on the Grid environment, are presented. Finally, our future research perspectives and the conclusions drawn are discussed.
2 Methodology

2.1 Basic Definition of the String Matching Problem with Bounded Gaps

The adoption of two uniformly Alphabets \([14] \Sigma_{DNA} = \{A, C, G, T\} \) and \(\Sigma_{Protein} = \{A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V\} \) is needed in order to define the string pattern matching problem with bounded gaps. Each letter is standing for the first letter of the chemical name of the nucleotide in a DNA sequence and each letter in the protein alphabet represents the amino acid abbreviation for each protein in a protein sequence.

Let \(X\) be a string drawn from \(\Sigma_{DNA}\). \(X\) represents an array \(X[1..n]\) of \(n\geq0\) symbols, where \(n=\text{length}(X)\) denotes the length of the string \(X\). \(X[i]\) denotes the \(i\)th symbol of \(X\). In addition, \(X[i..j]\) denotes the substring of \(X\) contained between \(i\)th and the \(j\)th symbol. Given a text \(T\) of length \(n\) and a pattern \(P\) of length \(m\), an occurrence with \(b\)-bounded gaps of \(P\) in \(T\) is an increasing sequence of indices \(i_1, i_2, \ldots, i_m\) such that (i) \(1 \leq i_1\) and \(i_m = i \leq n\) and (ii) \(i_{h+1} - i_h \leq b + 1\), for \(h = 1, 2, \ldots, m-1\). \(P \preceq^b T\) means that \(P\) has an occurrence with \(b\)-bounded gaps that terminates at position \(i\) in text \(T\). In the same way, an occurrence with \(a\)-bounded gaps of \(P\) in \(T\) is an increasing sequence of indices \(i_1, i_2, \ldots, i_m\) such that (i) \(1 \leq i_1\) and \(i_m = i \leq n\) and (ii) \(i_{h+1} - i_h \geq a + 1\), for \(h = 1, 2, \ldots, m-1\). \(P \preceq^a T\) means that \(P\) has an occurrence with \(a\)-bounded gaps that terminates at position \(i\) in text \(T\). Finally, an occurrence with \((a, b)\)-bounded gaps of \(P\) in \(T\) is an increasing sequence of indices \(i_1, i_2, \ldots, i_m\) such that (i) \(1 \leq i_1\) and \(i_m = i \leq n\) and (ii) \(a + 1 \leq i_{h+1} - i_h \leq b + 1\), for \(h = 1, 2, \ldots, m-1\). \(P \preceq^a_{(a,b)} T\) means that \(P\) has an occurrence with \((a, b)\)-bounded gaps that terminates at position \(i\) in text \(T\).

In [14], four algorithms are described. The interface for the Grid implements specifically the string pattern matching problem with \(b\) - bounded gaps, \(a\)-bounded gaps and \((a, b)\)-bounded gaps. The three implemented algorithms for the solution of the string pattern matching problems with bounded gaps are described as follows:

Problem 1 (upper bounded gaps) Given a text \(T\) of length \(n\), a pattern \(P\) of length \(m\) and a positive integer \(b\), the STRING PATTERN MATCHING PROBLEM WITH \(b\)-BOUNDED GAPS is to find all positions \(j\) in \(T\) such that \(P \preceq^b T\), for \(1 \leq j \leq n\).

Problem 2 (lower bounded gaps) Given a text \(T\) of length \(n\), a pattern \(P\) of length \(m\) and a positive integer \(a\), the STRING PATTERN MATCHING PROBLEM WITH \(a\)-BOUNDED GAPS is to find all positions \(j\) in \(T\) such that \(P \preceq^a T\), for \(1 \leq j \leq n\).

Problem 3 (lower & upper bounded gaps) Given a text \(T\) of length \(n\), a pattern \(P\) of length \(m\) and positive integers \((a, b)\), the STRING PATTERN MATCHING PROBLEM WITH \((a, b)\)-BOUNDED GAPS is to find all positions \(j\) in \(T\) such that \(P \preceq^a_{(a,b)} T\), for \(1 \leq j \leq n\).
In addition, the fourth algorithm mentioned in [14] is not implemented due to its nature of expecting many parameter, that is, \( a \) and/or \( b \) restrictions on every possible gap between every nucleotide. The specific interface could be expanded with Regular Expressions or with the help of a special parser for the input of the fourth algorithm. An other feature Regular Expressions can provide, is finding occurrences of patterns with bounded gaps using the IUPAC nucleotide code.

In other words, from a biologist’s point of view, Problems 1, 2 and 3 find all positions in a DNA or protein sequence \( T \) with at most \( b \) gaps, at least \( a \) gaps and between \( a \) and \( b \) gaps respectively, between every two nucleotides in the pattern specified. Setting \( a=0 \) and \( b=0 \) turns the problem in finding a specific pattern in a DNA sequence. The problems described above need plenty of time to run with large DNA sequences in length. For this reason, two complementary options are examined: Grid computing which offers the ability to run computationally and storage intensive applications [15] and a GUI friendly enough, in order to submit multiple instances of the algorithm with different \( a \) and/or \( b \). As a result, a biologist can benefit from the fact of retrieving his/her result in a reasonable amount of time.

### 2.2 The Implementation of the User Interface

The GUI in Fig. 1 incorporates all the necessary steps for submitting and managing a job in a Grid environment. The control panel (Submit job/s, Status, Create Proxy, Save Job/s, Load Job/s, Cancel Job/s) which offers a user-friendly job management, also exists in the specific User Interface designed for string pattern matching with Bounded gaps.

EGEE was the Grid infrastructure that our interface utilized and gLite 3.1 [16] was the necessary middleware for accessing the Grid platform. The GUI was developed in Java and WMProxy API (Application Programming Interface) was used for submitting, cancelling and retrieving the output. WMProxy is implemented as a web service. A web service allows us to take advantage of the benefits of the web, not only to provide information, but also to offer services to a greater community of possible users [17]. Within the bioinformatics community, an average end-user might need to access and use hundreds of databases and tools on a given day [18].

The creation of a VOMS (Virtual Organization Membership Service) proxy certificate for accessing the Grid and the status of the submitted jobs are handled by the gLite user interface via the use of java. For data uploading the gsiFTP client [11] was used for uploading the C++ implementation of the string pattern matching with bounded gaps algorithm.

All the scripts and the JDL (Job Description Language) files that are needed for the submission of the jobs are generated automatically via the GUI. This automation saves time and makes the submission of jobs for the naive user simpler enough. The only parameters the user has to fill in, are the number of the jobs, if he wants to perform a parameter study and then the names of the error and output files.
Fig. 1 The specific User Interface, SearchDNA Gaps that implements and eases the string pattern matching with bounded Gaps

a list box, the user has also the ability to choose from a list of predefined CEs (Computing Element), or a random one.

The uploading of the DNA or protein sequences in fasta format is performed, via the use of the GUI. It has to be mentioned that each fasta file should contain only one DNA sequence as an input, because in this implementation different patterns are utilized and searched for occurrences in the DNA sequence. The user is enabled to choose one of the three problems for string pattern matching with bounded gaps mentioned in Sect. 2.1.

An another important element of the specific user interface, is the submission and the management of multiple jobs (Status, Save, Load, Cancel, Get output). As a result, the parameter study for different values of the string pattern matching algorithm with bounded gaps \((a, b, a \text{ and } b)\) can be performed and retrieval of output is faster and efficient.

3 Test Case and Results

Our algorithm implementation was tested on the Grid EGEE platform with different DNA sequences that differ in size, downloaded randomly from NCBI. Figure 2\(^1\), illustrates these results. Twelve different DNA sequences ranged from 3626 to

\(^1\) Arabidopsis2 and Arabidopsis3 are segments of the Arabidopsis DNA sequence
30,423,563 base pairs and one small Pattern, ATGCSCG, were used as an input. For each DNA sequence, $a$, $b$, and $(a,b)$ parameters were initialized with the same values five times. The values of $a$, $b$ were chosen randomly, according to the three problems described in Sect. 2.1. The results are illustrated in Fig. 2 and one can easily draw the conclusion that the execution time of a job for a string pattern matching with bounded gaps depends on the length of the DNA sequences. More precisely, by keeping the pattern unchanged, if the length of a DNA sequence gets over the eighth million base pairs, then the execution time grows rapidly. For this reason, Grid can be used with multiple job submission and different parameter values, with large DNA sequences in length. Also, the small deviation in execution times for different $a$ and/or $b$ is due to the fact that CEs were chosen randomly.

![Fig. 2 Execution Times in seconds of the string pattern matching algorithm with bounded gaps for different DNA sequences and $(a, b, a$ and $b)$ parameters](image)

As far as Protein sequences concerned, it has to be noted that they are rather small in length. Titin protein with gi number (genInfo identifier), gi|108861911 was used for testing. The results produced were as they were expected to be. Due to the small sizes of protein sequences, the execution times are rather small, only five seconds. Today, personal computers are very fast and they can handle small DNA or protein sequences efficiently as Fig. 2 shows. The use of Grid is strongly recommended for large sequences, that require hours for the execution of a simple job of string pattern matching.
4 Discussion - Conclusions

In this paper, a specific user-friendly interface was developed for the execution of specific jobs, those for string pattern matching with bounded gaps in DNA or Protein sequences. The rationale of this work is to present the Grid infrastructure as a mean, that can execute computationally and storage exhaustive applications, in the concept of a parameter study approach, by using the multiple job submission feature of the proposed specific GUI. Table 1 summarizes and gives a brief explanation of each of the technical terms, projects, and resources appeared in the paper.

The applicability and the potentiality of the proposed specific GUI was illustrated by testing it with different DNA sequences that vary in the size of length. Protein sequences are rather small in size and the algorithm needs only seconds to be executed, so personal computers can be used for this kind of string pattern matching.

The implementation of the algorithm takes as input only one sequence at each job and examines for bounded gaps. In the future, the algorithm and the Grid interface will be changed to query a large number of different sequences saved in fasta format, at each job. Additionally, the combination of the string pattern matching with bounded gaps algorithm, with approximate string matching algorithms [19], will allow to find the positions of a text where a given pattern occurs, allowing a number of “errors” in the matches with bounded gaps. The parallelization of the algorithm with MPI (Message Passing Interface) and the use of multiple submission or job workflows could expand the features of the GUI, facilitating the user to submit fewer jobs.

The current implementation requires the installation of the gLite middleware in the computer hosting the GUI. This limitation is expected to be solved in the near future and we working on that. Finally, the need of a mobile Graphical User Interface for the Grid using the Java technologies, is in our nearest expectations and we are looking to present one in the nearest future.
<table>
<thead>
<tr>
<th><strong>Term</strong></th>
<th><strong>Brief Explanation</strong></th>
<th><strong>URL resource</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>gLite WMS</td>
<td>gLite Workload Management System</td>
<td><a href="http://glite.web.cern.ch/glite/packages/R3.1/deployment/glite-WMS/glite-WMS.asp">http://glite.web.cern.ch/glite/packages/R3.1/deployment/glite-WMS/glite-WMS.asp</a></td>
</tr>
<tr>
<td>gLite UI</td>
<td>gLite User Interface</td>
<td><a href="http://glite.web.cern.ch/glite/packages/R3.1/deployment/glite-UI/glite-UI.asp">http://glite.web.cern.ch/glite/packages/R3.1/deployment/glite-UI/glite-UI.asp</a></td>
</tr>
<tr>
<td>EGEE</td>
<td>Enabling Grids for E-sciencE</td>
<td><a href="http://www.eu-egee.org/">http://www.eu-egee.org/</a></td>
</tr>
<tr>
<td>NGS</td>
<td>The National Grid Service</td>
<td><a href="http://www.grid-support.ac.uk/">http://www.grid-support.ac.uk/</a></td>
</tr>
<tr>
<td>HealthGrid</td>
<td>HealthGrid Community</td>
<td><a href="http://community.healthgrid.org/">http://community.healthgrid.org/</a></td>
</tr>
<tr>
<td>BioinfoGRID</td>
<td>The BioinfoGRID project</td>
<td><a href="http://www.bioinfogrid.eu/">http://www.bioinfogrid.eu/</a></td>
</tr>
<tr>
<td>g-Eclipse</td>
<td>g-Eclipse project</td>
<td><a href="http://www.geclipse.org/">http://www.geclipse.org/</a></td>
</tr>
<tr>
<td>GuiGen</td>
<td>GuiGen is a comprehensive set of tools for creating customized graphical user interfaces (GUIs)</td>
<td><a href="http://www.zib.de/schintke/guigen/index.en.html">http://www.zib.de/schintke/guigen/index.en.html</a></td>
</tr>
<tr>
<td>P-GRADE</td>
<td>Parallel Grid Run-time and Application Development Environment</td>
<td><a href="http://www.p-grade.hu/">http://www.p-grade.hu/</a></td>
</tr>
<tr>
<td>Globus Toolkit</td>
<td>The Globus Toolkit is an open source software toolkit used for building Grid systems and applications</td>
<td><a href="http://www.globus.org/">http://www.globus.org/</a></td>
</tr>
<tr>
<td>Condor Project</td>
<td>The Condor Project</td>
<td><a href="http://www.cs.wisc.edu/condor/">http://www.cs.wisc.edu/condor/</a></td>
</tr>
<tr>
<td>CBG algorithms</td>
<td>Classes of characters and Bounded size Gaps algorithms</td>
<td>Navarro G, Raffinot M (2001)</td>
</tr>
<tr>
<td>Regular Expressions</td>
<td>Regexp (for short) is a special text string for describing a search pattern</td>
<td><a href="http://java.sun.com/docs/books/tutorial/essential/regex/">http://java.sun.com/docs/books/tutorial/essential/regex/</a></td>
</tr>
<tr>
<td>GUI</td>
<td>Graphical User Interface</td>
<td><a href="http://infovis.cs.vt.edu/GUI/java/">http://infovis.cs.vt.edu/GUI/java/</a></td>
</tr>
<tr>
<td>WMProxy</td>
<td>WMProxy is a new component to access the gLite Workload Management System (WMS)</td>
<td><a href="http://trinity.datamat.it/projects/EGEE/wiki/wiki.php?n=WMProxyService">http://trinity.datamat.it/projects/EGEE/wiki/wiki.php?n=WMProxyService</a></td>
</tr>
<tr>
<td>gsiFTP client</td>
<td>A client developed by using GridFTP (GSI enabled FTP) protocol for the file transfers</td>
<td><a href="http://www-unix.globus.org/cog/distribution/1.1/API.html">http://www-unix.globus.org/cog/distribution/1.1/API.html</a></td>
</tr>
</tbody>
</table>
References

1. Baxevanis A. D. , Francis Ouellette B. F.: Bioinformatics and the Internet. In: BIOINFOR-
7. g-Eclipse at http://www.geclipse.eu/
16. gLite documentation at http://glite.web.cern.ch/glite/documentation/
17. Avellino G. et al, 2006. Flexible job submission using web services: The gLite WMProxy experience. 15th International Conference on Computing In High Energy and Nuclear Physics, Mumbai, India, pp.831 - 835