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

Lefteris Koumakis, Stelios Sfakianakis, Vassilis Moustakis, and George Potamias

Institute of Computer Science, FORTH

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 disease 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 Grid*s. 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 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-depended 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.

```
<experiment total-assays="78" total-samples="78" total="1" revision="080925" ver-
sion="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 genotype 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 Servic-

es (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 grand permissions to user's account in the Grid and services repository give access to available services. Then the user composes and draws he desired workflow. At the next step the editor translates (or compile) 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".

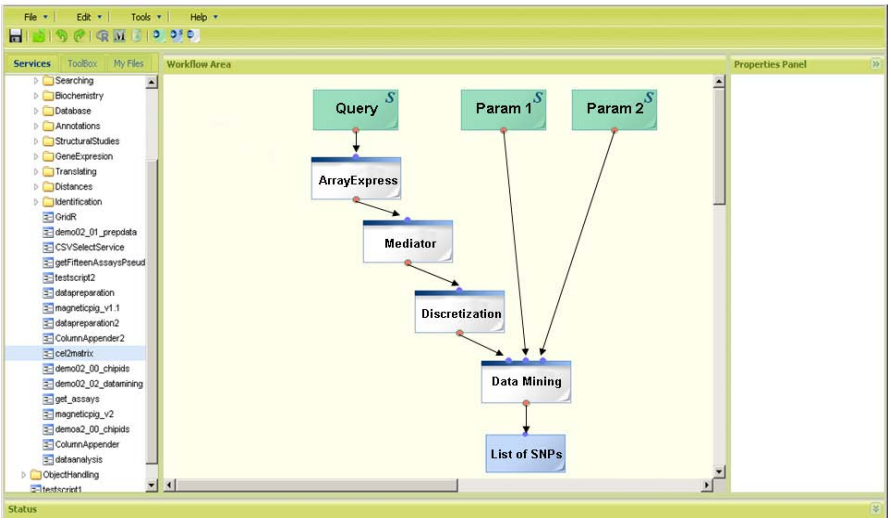


Fig. 3.2. The GG2P scientific workflow as implemented in ACGT's Workflow Editor and Enactment Environment (WEEE). Web services include: ArrayExpress, Mediator, Discretization, and Data Mining. Services are activated by a Query (top part). Deployment of Data Mining also needs specification of parameters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

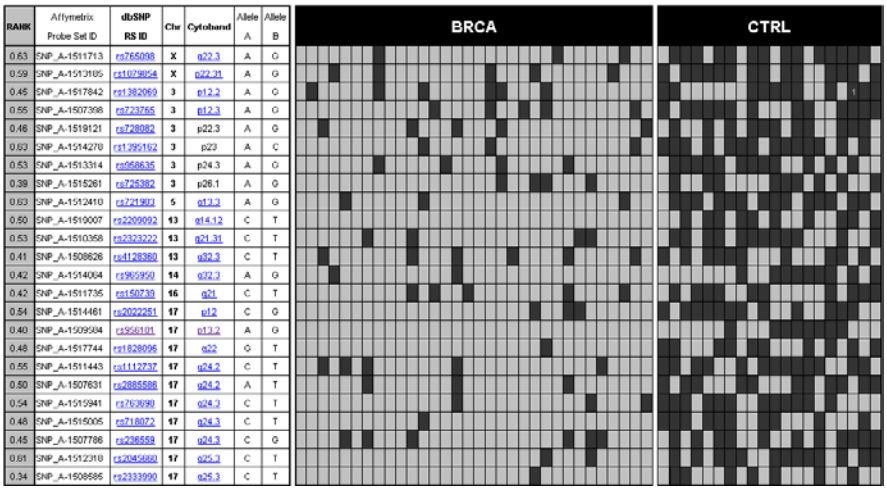
The Affymetrix SNP genotyping platforms produce processed data files where, each SNP receives three different values: AA and BB that represent paternal or maternal homozygosity statuses, respectively, and AB for heterozygosity ones. The '0' and '1' nominal values are assigned to the AA/BB and AB SNP 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 offsping's functional allele is inactivated by the mutation. In such situations, normal tumor suppressor functionality is inactivated and tumorigenesis events are almost certain.



RANK	Affymetrix Probe Set ID	dbSNP RS ID	Chr	Cytoband	Allele A	Allele B
0.63	SNP_A-1511713	rs765098	X	q22.3	A	G
0.59	SNP_A-1513105	rs1079854	X	q22.31	A	G
0.45	SNP_A-1517842	rs1382069	3	p12.2	A	G
0.55	SNP_A-1507398	rs723765	3	p12.3	A	G
0.46	SNP_A-1519121	rs728082	3	p22.3	A	G
0.60	SNP_A-1514278	rs1395162	3	p23	A	C
0.53	SNP_A-1513314	rs968635	3	p24.3	A	G
0.39	SNP_A-1515261	rs725382	3	p26.1	A	G
0.60	SNP_A-1512410	rs721903	5	q13.3	A	G
0.50	SNP_A-1519007	rs2209092	13	q14.12	C	T
0.53	SNP_A-1510358	rs2323222	13	q21.31	C	T
0.41	SNP_A-1508626	rs4128380	13	q32.3	C	T
0.42	SNP_A-1514064	rs1965950	14	q32.3	A	G
0.42	SNP_A-1511735	rs150730	16	q21	C	T
0.54	SNP_A-1514461	rs2022251	17	p12	C	G
0.40	SNP_A-1509584	rs956101	17	p13.2	A	G
0.48	SNP_A-1517744	rs1828096	17	q22	G	T
0.55	SNP_A-1511443	rs1112737	17	q24.2	C	T
0.50	SNP_A-1507631	rs2885588	17	q24.2	A	T
0.54	SNP_A-1515941	rs763690	17	q24.3	C	T
0.48	SNP_A-1515005	rs718072	17	q24.3	C	T
0.45	SNP_A-1507786	rs236559	17	q24.3	C	G
0.61	SNP_A-1512318	rs2045880	17	q25.3	C	T
0.34	SNP_A-1508686	rs2333990	17	q25.3	C	T

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 1nd 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 heterozigosity patterns, a common situation in tumorgenesis. 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

Arkin A et al (Eds.) (2005) Web services business process execution language. Version 2.0. http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html Accessed 8 March 2009

Cannataro M, Talia D (2003) The Knowledge Grid. Communications of the ACM 46(1):89 –93, 2003

Curbera F et al (2002) Unraveling the web services web: An Introduction to SOAP, WSDL, and UDDI. IEEE Internet Computing 6(2):86–93

De Roure D, Jennings NR, Shadbolt NR (2005) The Semantic Grid: Past, Present, and Future. Proceedings of the IEEE 93(3):669–681

Deelman E, Zhao Z, Belloum A (Eds.) (2006) Scientific Programming Journal, special issue on workflows to support large-scale science 14(3–4)
Ellsworth EE et al (2003) High-Throughput Loss of Heterozygosity Mapping in 26 Commonly Deleted Regions in Breast Cancer. Cancer Epidemiology Biomarkers & Prevention 12:915–919
Foster I (2003) The Grid: Computing without bounds. Scientific American 288(4):60–67.
Fox G, Gannon D. (Eds.) (2006) Concurrency and Computation: Practice and Experience, Special Issue on Workflow in Grid Systems 18(10)
Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. JMLR: Special Issue on Variable and Feature Selection 3:1157–1182
Mao X et al (2005) Genetic losses in breast cancer: toward an integrated molecular cytogenetic map. Cancer Genetics and Cytogenetics 160(2):141–151
Nunkesser R, Bernholt T, Schwender H, et al (2007) Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. Bioinformatics 23(24): 3280–3288
Potamias G, Koumakis L, Moustakis VM (2005) Enhancing Web Based Services by Coupling Document Classification with User Profile. International Conference on Computer as a Tool (EURCON 2005) 1:205–208
Potamias G, May M. Ruping S (2006) Grid-based Knowledge Discovery in Clinico-Genomic Data. Lecture Notes in Bioinformatics (LNBI) 4345:219–230
Potamias G., Koumakis L, Moustakis V (2004) Gene Selection via Discretized Gene Expression Profiles and Greedy Feature-Elimination. Lecture Notes in Artificial Intelligence (LNAI) 3025:256–266, (2004)
Pukacki J et al (2006) Programming Grid Applications with Gridge, Computational Methods in Science and Technology 12(1):47–68
Richardson A et al (2006) X chromosomal abnormalities in basal-like human breast cancer. Cancer Cell 9(2):121–32
Schwender H, Ickstadt K, Rahnenführer J (2008) Classification with high-dimensional genetic data: assigning patients and genetic features to known classes. Biometrical Journal 50(6):91 – 926
Seitz S et al (2001) Detailed deletion mapping in sporadic breast cancer at chromosomal region 17p13 distal to the TP53 gene: association with clinicopathological parameters. Journal of pathology 194(3):318–326
Sfakianakis S, et al (2009) Web-based Authoring and Secure Enactment of Bioinformatics Workflows. 4th International Workshop on Workflow Management (ICWM2009), Geneva, Switzerland
Shen C-Y et al (2000) Genome-wide Search for Loss of Heterozygosity Using Laser Capture Microdissected Tissue of Breast Carcinoma: An Implication for Mutator Phenotype and Breast Cancer Pathogenesis. Cancer Research 60:3884–3892
Singh MP, Vouk MA (1997) Scientific workflows: Scientific computing meets transactional workflows.
http://people.engr.ncsu.edu/mpsingh/papers/databases/workflows/sciworkflows.html Accessed 8 March 2009
Witten IH, Frank E (2005) Data Mining: Practical machine learning tools and techniques (2nd Edition), Morgan Kaufmann, San Francisco
Zhou N, Wang L (2007) Effective selection of informative SNPs and classification on the HapMap genotype data. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid= 2245981. Accessed 8 march 2009
Zhuge H (2004) The Knowledge Grid. World Scientific Singapore