31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Supporting Genotype-to-Phenotype Association Studies with Grid-enabled Knowledge Discovery Workflows

Lefteris Koumakis, Vassilis Moustakis, Manolis Tsiknakis, Dimitris Kafetzopoulos, and George Potamias

Abstract—**Web Services and Grid-enabled scientific workflows are of paramount importance for the realization of efficient and secure knowledge discovery scenarios. This paper presents a Grid-enabled Genotype-to-Phenotype discovery scenario (GG2P), which is realized by a respective scientific workflow. GG2P supports the seamless integration of SNP genotype data sources, and the discovery of indicative and predictive genotype-to-phenotype association models – all wrapped around custom-made Web Services. GG2P is applied on a whole-genome SNP-genotyping experiment (breast cancer vs. normal/control phenotypes). A set of about 100 indicative SNPs are induced with very high classification performance. The biological relevance of the findings is supported by the relevant literature.**

I. Introduction

With the completion of the Human Genome Project and the entrance to the post-genomics era, associated technology developments have accelerated the process of analyzing entire genomes. In turn this has catalyzed the major development of predictive, preventive and personalized medicine, which will impact on clinical practice. In particular, it has provided access to the extensive human genome variability in the form of SNPs (single nucleotide polymorphisms), some of which predispose to disease. This knowledge introduces the prospect of clinical prognosis based on identification of susceptibility genes. It is likely that a predictive medicine will gradually emerge, capable of determining a probabilistic 'future health history' for each individual. As individuals maintain unique genotype information, inter-individual genome variation plays a major role in differential development and disease processes. Background genetic effects (modifier genes), epitasis, somatic variation, and environmental factors all complicate the situation. Strategies do, however, now exist to study complex disease genetics. The raised needs concern on one hand, the creation and effective integration of ever-larger datasets, and on the other, the processing of these datasets to induce reliable knowledge. At present, too little is known about which SNPs to type for complex and multi-factorial diseases, because varying linkage disequilibrium patterns in different populations make it unsuitable for unsorted patient samples. In this diverse and semantically heterogeneous clinico-genetics environment, clinical data-processing and decision-making processes become more demanding in terms of their domain of reference, i.e., population-oriented and lifetime *clinical profiles enriched by evidential genomic/genetic information*. Moreover, health prevention- a key-issue for reliable and effective medical care, and the related epidemiological studies become more dependent on information transfer and exchange. The ability of healthcare professionals to be informed, to consider and to adapt fast to the potential changes and advances of the medical practice is of crucial importance for future decision-making. Therefore, methodologies, systems and tools to extend the capacity of the *unaided mind* are required, to couple the details of knowledge about a problem with the relevant knowledge from combined evidenced clinical and genomic knowledge repositories. The use of genetic data in addition to clinical symptoms for medical decision-making will contribute to the expected, and continued shift towards *evidence-based clinico-genomic medicine*. Such a visionary objective can only be realized with an enormous investment into: (i) the creation of standardized databases that combine clinical history, symptoms and signs, laboratory and procedural results and genetic data in raw and processed formats; and (ii) the extraction of knowledge out of the respective databases, their expert interpretation and matching against existing computational models; and (iii) the incorporation such knowledge into standardized clinical guidelines where feasible.

Knowledge Discovery and Data Mining are the most prominent approaches to automated 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.

In this paper we present an integrated scenario to support seamless access and analysis of 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 disease phenotypes. The employed data-mining technique is founded on a novel feature selection algorithm. The whole approach is realized in a Grid-enabled scientific workflow editor and enactment environment, and presents an integrated scenario aiming to support Grid-enabled Genotype-to-Phenotype (GG2P) association studies. In particular, GG2P induce discriminant and predictive SNP-phenotype association models linking the results with Ensembl, a state-of-the-art genome browser.

Manuscript received April 7, 2009. This work was supported in part by the EU ACGT (FP6-ICT-2005-026996) and GEN2PHEN (FP7-HEALTH-2007-200754) integrated projects. L. Koumakis, M. Tsiknakis and G. Potamias (corresponding author) are with the Institute of Computer Science, FORTH, N. Plastira 100, 70013 Heraklion, Crete, Greece (phone: 30-2810-391693; fax: 30-2810-39160; e-mails: {koumakis, tsiknaki, potamias}@ics.foth.gr). V. Moustakis is with the Institute of Computer Science, FORTH and with the Dept. of Production Engineering & Management, Chania, Crete,. (e-mail: moustaki@ics.forth.gr). D. Kafetzopoulos is with the Institute of Molecular Biology & Biotechnology, FORTH, N. Plastira 100, 73100 Heraklion, Crete, Greece (phone: 30-2810-391594; fax: 30-2810-391101; e-mail: kafetzo@imbb.forth.gr).

II. ENABLING TECHNOLOGY

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. 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 (G2P) 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. Web Services standards present the most popular and successful integration methodology approach. Standards support 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. Workflow Management Coalition 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. A Workflow Management System (WMS) 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). 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. 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. 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. WEEE is based on the BPEL (Arkin et al 2005) workflow standard and supports the BPEL representation of complex bioinformatics workflows. In WEEE a generic data protection framework is used, which is based on a technical security infrastructure as well as on organizational measures (Claerhout et al., 2008). The ACGT Grid environment is supported by the Gridge Globus-comliant toolkit (www.gridge.org; www.globus.org; Pukacki et al., 2006).

III. 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.

A. Data access and retrieval

Using Web Services from the European Bioinformatics Institute's (EBI) repository (http://www.ebi.ac.uk/Tools/webservices/). Access was facilitated by ArrayExpress web service. The dataset used in this study refers to a genotyping experiment of 78 sample hybridizations performed on the Affymetrix GeneChip Human Mapping 10K Array Xba 131 (Mapping10K_Xba131 11560 SNPs) 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.

B. Data mediation

The response of ArrayExpress web service is an XML file with links to phenotypic ('sdrf') and genotype ('fgem') tags in the file. We utilized a special parser to extract the needed information from the XML file. 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.

C. 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 is performed, the 'filtered' dataset is transformed into the ARFF format - a de facto standard for machine learning, and supported by the Weka machine learning package (http://www.cs.waikato.ac.nz/ml/weka/) (Witten and Frank 2005).

D. 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, 2006).

IV. GG2P IN ACTION

For the realization of GG2P scenario we used part of the ACGT Grid infrastructure – the Data Management System (DMS), the service repository and the WEEE workflow editing and execution environment. The 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 ACGT WEEE workflow editor is a Web2 BPEL-compliant application installed in a Grid node. The first WEEE/GG2P web service takes as input a query and returns an XML file with information about all the related to the query experiments in the EBI ArrayExpress repository (query: "homo sapiens" & "breast cancer" & "genotype" & "affymetrix" &

"Mapping10K_Xba131"). 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 selected SNPs. Finally, an html file is formed and is used for the visualization of results.

V. Results and Discussion

Affymetrix SNP genotyping platforms produce processed data files where, each SNP receives three different values: AA and BB that represent paternal or maternal homozygosity genotypes, 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.

Figure 1 visualizes just the top 24 of the most discriminant SNPs (the ones with the highest ranks) 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 3 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') genotypes, respectively.

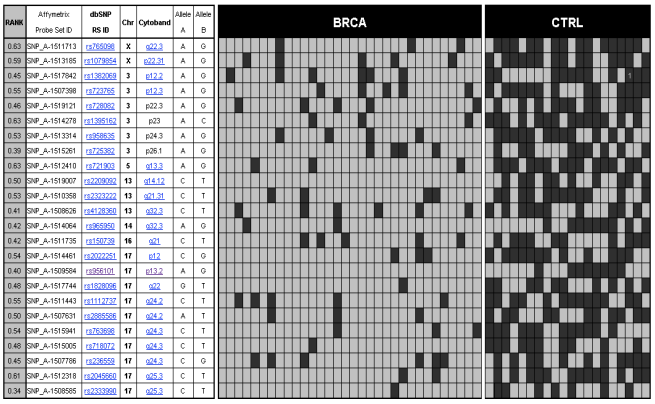

Fig. 1. The induced most discriminant and highest ranked BRCA vs. CTRL SNPs (for the ArrayExpress E-GEOD-3743 genotyping experiment) – gray shaded and dark shaded cells indicate homozygosity and heterozygocity genotypes, respectively. It can be easily observed that LOH (Loss Of Heterozygosity) patterns dominate the BRCA cases

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.

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 reported 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).

VI. 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 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. Results presented herein are cast in the same realm with work reported by Trégoüet et al (Trégoüet et al 2009) in coronary heart disease and demonstrate that grid technology coupled with workflow modeling and web services provide an effective team formation toward SNP and G2P discovery.

Our future R&D plans 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).

Acknowledgement

This work is partially supported by the European Commission's Sixth and Seventh Framework Program in context of GEN2PHEN (FP7-HEALTH-2007-200754) and ACGT (FP6-ICT-2005-026996) projects and the ACTION-Grid International Cooperative Action, FP7-ICT-2007-2.

References

[1] A. Arkin, et al., (2007, April 11). Web services business process execution language (Version 2.0) [Online]. Available: http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html

[2] M. Cannataro, and D. Talia, "The Knowledge Grid," *Communications of the ACM* , vol. 46, no. 1, pp. 89 –93, 2003

[3] B. Claerhout, N. Forgo, T. Krugel, M. Arning, and G. De Moor, "A Data Protection Framework for Transeuropean genetic research projects." *Studies in health technology and informatics*, vol. 141, p. 67, 2008.

[4] F. Curbera, et al., "Unraveling the web services web: An Introduction to SOAP, WSDL, and UDDI," *IEEE Internet Computing*, vol. 6, no. 2, pp. 86–93, 2002.

[5] D. De Roure, N.R. Jennings, and N.R. Shadbolt, "The Semantic Grid: Past, Present, and Future," *Proceedings of the IEEE*, vol. 93, no. 3, pp. 669–681, 2005.

[6] E. Deelman, Z. Zhao, and A. Belloum, eds., *Scientific Programming Journal*, special issue on workflows to support large-scale science vol. 14, no. 3–4, 2006.

[7] E.E. Ellsworth, et al., "High-Throughput Loss of Heterozygosity Mapping in 26 Commonly Deleted Regions in Breast Cancer," *Cancer Epidemiology Biomarkers & Prevention*, vol. 12, pp. 915–919, 2003.

[8] I. Foster, "The Grid: Computing without bounds," *Scientific American*, vol. 288, no. 4, pp. 60–67, 2003.

[9] G. Fox, and D. Gannon, eds., *Concurrency and Computation: Practice and Experience*, special issue on Workflow in Grid Systems, vol. 18, no. 10, 2006.

[10] I. Guyon, and A. Elisseeff, "An introduction to variable and feature selection," *JMLR*: Special Issue on Variable and Feature Selection, vol. 3, pp. 1157–1182, 2003.

[11] X. Mao, et al., "Genetic losses in breast cancer: toward an integrated molecular cytogenetic map," *Cancer Genetics and Cytogenetics*, vol. 160, no. 2, pp. 141–151, 2005.

[12] R. Nunkesser, et al., "Detecting high-order interactions of single nucleotide polymorphisms using genetic programming," *Bioinformatics*, vol. 23, no. 24, pp. 3280–3288, 2007.

[13] G. Potamias, L. Koumakis, and V. Moustakis, "Enhancing Web Based Services by Coupling Document Classification with User Profile," in *Proc. International Conference on Computer as a Tool* (EUROCON 2005), 2005, vol. 1, pp. 205–208.

[14] G. Potamias, M. May, and S. Ruping, "Grid-based Knowledge Discovery in Clinico-Genomic Data," *Lecture Notes in Bioinformatics* (LNBI), vol. 4345, pp. 219–230, 2006.

[15] G. Potamias, L. Koumakis, and V. Moustakis, "Gene Selection via Discretized Gene Expression Profiles and Greedy Feature-Elimination," *Lecture Notes in Artificial Intelligence* (LNAI), vol. 3025, pp. 256–266, 2004.

[16] J. Pukacki, et al., "Programming Grid Applications with Gridge," *Computational Methods in Science and Technology*, vol. 12, no. 1, pp. 47–68, 2006.

[17] A. Richardson, et al., "X chromosomal abnormalities in basal-like human breast cancer," *Cancer Cell*, vol. 9, no. 2, pp. 121–32, 2006.

[18] H. Schwender, K. Ickstadt, and J. Rahnenführer, "Classification with high-dimensional genetic data: assigning patients and genetic features to known classes," *Biometrical Journal*, vol. 50, no. 6, pp. 911–926, 2008.

[19] S. Seitz, et al., "Detailed deletion mapping in sporadic breast cancer at chromosomal region 17p13 distal to the TP53 gene: association with clinicopathological parameters," *Journal of pathology*, vol. 194, no. 3, pp. 318–326, 2001.

[20] S. Sfakianakis, et al., "Web-based Authoring and Secure Enactment of Bioinformatics Workflows," *4th International Workshop on Workflow Management* (ICWM2009), Geneva, Switzerland, 2009

[21] G. Potamias, L. Koumakis, and V. Moustakis, "Enhancing Web Based Services by Coupling Document Classification with User Profile," in *Proc. International Conference on Computer as a Tool* (EUROCON 2005), 2005, vol. 1, pp. 205–208.

[22] C-Y. Shen, et al., "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, vol. 60, pp. 3884–3892, 2000.

[23] M.P. Singh, and M.A. Vouk. Scientific workflows: Scientific computing meets transactional workflows [Online]. Available: http://people.engr.ncsu.edu/mpsingh/papers/databases/workflows/sciworkflows.html Accessed 8 March 2009

[24] D-A. Trégoüet et al, "Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease", *Nature Genetics*, 41(3): 283-285, 2009.

[25] I.H. Witten, and E. Frank, *Data Mining: Practical machine learning tools and techniques* (2nd Edition). San Francisco: Morgan Kaufmann, 2005.

[26] N. Zhou, and L. Wang, "Effective selection of informative SNPs and classification on the HapMap genotype data," *BMC Bioinformatics*, 8: 484, 2007.

[27] H. Zhuge, *The Knowledge Grid*. Singapore: World Scientific, 2004.