MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-1304

Scientific discovery workflows in bioinformatics: A scenario for the coupling of molecular regulatory pathways and gene-expression profiles

Alexandros Kanterakis, Giorgos Potamias, Giorgos Zacharioudakis, Lefteris Koumakis, Stelios Sfakianakis, Manolis Tsiknakis

Foundation for Research and Technology-Hellas, Institute of Computer Science (FORTH-ICS), Heraklion, Crete, Greece

Abstract

Scientific workflow technologies and tools have become an important weapon in the arsenal of the bioinformaticians and computational biologists. To support this view we present a typical exploratory data analysis scenario involving the combination of information from Gene Regulatory Networks and gene expression data. We further describe the implementation of this scenario using the Workflow Environment implemented in the context of a large EU funded project. In this process desirable features that similar environments should offer are identified and analyzed. The ICT platform presented is evaluated using the chosen scenario as a benchmark. Finally we conclude with an outlook to future work.

Keywords:

Bioinformatics, Semantics, Grid, Web services, Scientific workflows

Introduction

In the new era originated with the successful completion of the human genome sequencing projects molecular biology research has moved from the experimental laboratory bench to systems biology approaches enabled by the *in-silico* (computer-based) study of the huge biological data sets produced by the high throughput technologies such as DNA microarrays [1,2]. To tackle these challenges some new computational tools and techniques have been developed to offer complex data analysis and visualization. Scientific workflows [3,4] present an "umbrella" term to describe the use of computational tools that help to automate data collection, analysis and processing tasks of scientific experiments.

This paper aims to present a complex, yet characteristic bioinformatics scenario and its implementation through the scientific workflow environment implemented in the context of an FP6 EU funded project with the acronym ACGT (Advancing Clinico-Genomic Clinical Trials on Cancer: Open Grid Services for improving Medical Knowledge Discovery, http://www.eu-acgt.org/). The objective of the ACGT project is to develop and test an ontology driven, semantic grid services infrastructure enabling efficient execution of discovery-driven analytical workflows in the context of multi-centric, post-genomic clinical trials [5].

An indicative bioinformatics scenario

One of the most significant sources of genomic functionality stems from Gene Regulatory Networks (GRNs). The compilation of GRNs known today was a complicated effort since it demands the combination and distillation of numerous and dispersed source of information originating mainly from the biomedical literature. As a side effect, devised GRNs are far from complete since, not all genetic interaction are known and the one that are already modeled have not been fully documented. Another issue concerns the fact that each GRN models a separate cellular functionality but this abstraction is rather obsolete. Today, while we move towards the realization of the systems biology vision, we tend to view genetic interactions as parts of a holistic and general mechanism that govern the cellular functionality. Therefore, we need to locate all the different interactions depicted in GRNs and to examine them out of the context of a specific and isolated cellular function (i.e. apoptosis, cell cycle, signaling etc) and move to a more unified and systemic point of views [6].

Decomposition

The first step of our approach is the decomposition of known and established GRNs. Using as our major source of GRNs the KEGG repository [7], we acquired all available regulatory networks related to cellular processes in their special KGML format (http://www.genome.jp/kegg/xml/). As regular XML files, the GRNs encoded through KGML can easily be parsed and modeled as a usual directed graph where, vaguely speaking, genes are the nodes and molecular interactions are the edges. Through the devise and utilization of a specific algorithm we can identify all possible paths contained in each graph. A path is a sequence of nodes and edges that could be followed in a graph in order to form a route from one node to another. Just a single presentation of a node or an edge is permitted in a path. In our effort to manage the paths as molecular functions, we assigned additional values to each gene according to its functional state in the path. We assume that each gene possess a particular functional role either by being in an "ON" or "OFF" state [8]. These values represent the possible activated/non-activated states or, expressed/non-

expressed levels of a gene during a certain cellular process. Moreover, these values are determined by the semantics of the interaction between two genes. Namely, the characterization of a gene in a path as being on an "ON" or "OFF" state is unambiguously specified by its preceding direct regulatory links (edges) with other genes. According to the semantics of the KGML file there are various types of gene interactions but since the functionality of many of them is similar we narrowed down to four basic different kinds of molecular interaction: 'activation', 'inhibition', 'association' and 'disassociation'. These interactions determine the "ON" / "OFF" state of a gene according to the following rules: If a gene is the first node of a path then it should be "ON". Otherwise, if its preceding relation is 'activation' then this gene should be "ON", and if it is 'inhibition' it should be "OFF". In case that the previous relationship is 'association' then the gene is considered to posses the same state as the previous gene in the path whereas, for a 'disassociation' relation, the gene is considered to posses the opposite state of its preceding gene in the path. The intuition that underlies these rules is that: if a gene is on an "ON" or "OFF" state there should be a reason for this. The linear reasoning that we provide goes opposite the direction of the path. As a generalization we could say that the state of a gene depends on two things: the state of its directly preceding genes in the path, and on the nature of its direct preceding interaction that govern its regulation. Thus, each path can be perceived as a history of events that explain the state of each gene.

Microarray Discretization

The set of paths with "ON"/"OFF" annotated genes comprise a functional decomposition of the available GRNs. Our aim is to verify the conditions under which these functions either "work" or "stay idle". This validation can be done by incorporating knowledge from microarray gene expression measurement over tissue samples pre-classified against two phenotypic categories, let "A" and "B". In order to validate the "ON"/"OFF" values of the genes in the paths we discretized the gene expression values into "High" and "Low" values [9, 10].

Combination

For each path Pi originating from the functional GRN decomposition we can determine if it is in an "operational" or in an "idle" (passive) state for any microarray sample Sj as follows: path Pi is considered to be operational for sample Sj, if (and only if) the genes that have "ON" values in Pi have "High" values in Sj, and respectively, the genes that have "OFF" values in Pi have "Low" values in Sj. If there is any misalignment between the "ON"/"OFF" values in the path and the respective "High"/"Low" values in the sample then, the path is considered to be idle for this sample. Applying this procedure for all samples then we can form a contingency table between the operational/idle states of a path and the "A" / "B" class phenotypes of the samples. Each entry in table represents the number of the microarray samples that belong to the "A" (or "B") class and are operational (or idle) in the path. By calculating the contingency table's Fisher's exact test figure we assess the significance of this association. Specifically, paths with low p (<0.001) indicate that they represent cellular functionalities that behaves differently between tissues belonging to class "A" and "B". Thus, by sorting all the paths in ascending p-value order, we gain insight to the cellular mechanisms that explain this differentiation.

Materials and Methods

It is common knowledge that the requirements for the management of the biological data are very demanding because of their size and complexity, quality properties (missing values or noisy data are frequent), and the inherent heterogeneity of the domain. These new requirements have given rise to modern software engineering methodologies and tools, such as the Grid [11] and the Web Services. These new technologies aim to provide the means for building sound and scalable data integration, management, and processing frameworks.

In order for these new technologies to become exploitable by the biologists and bioinformaticians a user friendly environment needs to be in place. This is an already recognized need, and a number of tools have been developed, such as the Taverna Workbench [12], Kepler [13], and Triana [14] to offer efficient and effective scientific workflow environments. The posted challenges related to the provision of state of art discovery processes in scientific workflows are well documented in [15]. In particular a major requirement is that the system should support the "reproducibility" of the results so that the same or other scientists can validate the whole process. Security and trust is another very important aspect of these environments, which means that the users trust the tools that will not inflict harm to their data or that their private data will be subject to unintended analysis by other users.

More specifically for the implementation of the scenario described above the following requirements should be satisfied:

- Secure controlled access to the user data. Although the general consensus is that the data should be shared in order to accelerate the scientific discovery process it is of course the right of the data curator not to share their data especially when sensitive patient information is the case. The microarray data used in the experiment are stored in the Grid Data Management System [16] in the user's private storage area and respective access rights require user's credentials.
- Access to "third party" publically available information sources. In particular the scenario requires the retrieval of the relevant gene regulatory networks from online stores such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [7].
- A number of processing steps need to be performed: decomposition of networks, discretization of microarray data, etc. For some of these tasks (e.g. discretization) there are available tools but for others the user should be able to enhance the functionality of the system by incorporating custom and reusable analytical methods and tools.

- The reliability and scalability of the system should guarantee that the end user would get the outcome of experiment in a logical time frame in spite of the complexity of the analyses or the size of the involved data sets.
- The system should be intuitive and reflect the user's way of thinking and assist to achieve his/her goals. Specifically, user friendliness and presentation of information at more domain specific and abstract user's conceptualization levels are important characteristics.
- Interoperability issues and heterogeneities among the tools and data sets should be resolved at the minimum cost and disturbance of the user. The use of metadata for the specification of additional aspects such as intent, policies, "meaning", etc., is a critical factor and ease to overcome such difficulties [17].

The ACGT Workflow Environment

To assist bioinformaticians in building their complex scientific workflows, a Workflow Editor and Enactment Environment, called WEEE [18], have been designed and implemented. They consist of a suite of graphical tools that allow a user to combine different web services into complex workflows. This environment is accessible through the ACGT Portal and therefore features a web based graphical user interface. It supports searching and browsing of a tool (service) registry and of respective data sources, as well as their orchestration' and composition through an intuitive and user friendly interface. The created scientific workflows can be stored in a user's specific area and later retrieved and edited so new versions of them can be produced. The designed workflows can be executed in a remote machine or even in a cluster of machines in the Grid so there is no burden imposed on the user's local machine since the majority of computation and data transfer of the intermediate results are taken place in the Grid where the services are run. The 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 [19] workflow standard and supports the BPEL representation of complex bioinformatics workflows.

In the following section we discuss the approach we have followed in ACGT in order to support these requirements using the scenario described in the introduction as a "yardstick".

Results

The workflow implementing the bioinformatics scenario described above is shown in Figure 1. The workflow consists of various web services, each implementing a functional unit of the whole workflow scenario. The core services (activities) are the DiscretizationService ("EntropyDiscretize"), the DecompositionService ("decompose") and the CombinationService ("combine") which independently implement the corresponding functions mentioned previously. Besides these core activities, additional entities are introduced to either make feasible the interconnections between the activities or to further enhance the functionality. In the former case the FileService ("writeFile") is used for data retrieval and storage on the Grid and in the latter case a Biomoby service [20] is used for visualizing the GRN as an image. We have used the BioMoby web service called "getKeggPathwayAsGif", which, given a KEGG network identifier, returns a GIF image of the corresponding pathway.

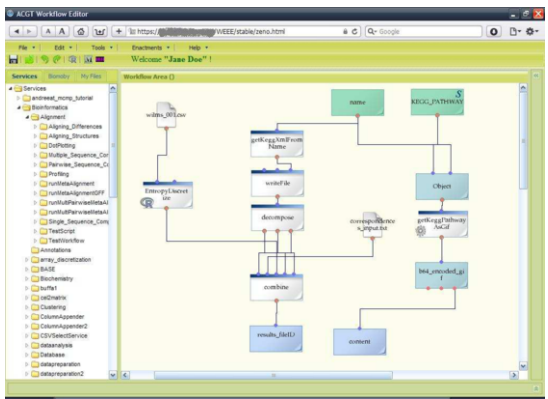

Figure 1 - The workflow implementing the coupling GRN/gene-expression scenario as realized in the ACGT Workflow Editor

The information is passed from one service to another by reference, using files on the Grid. This is done not only for performance reasons, but also for having a means to monitor and document the execution of the workflow, store intermediate results and re-use of data. Since BPEL supports the parallel execution of tasks, some of the services of the workflow are executed simultaneously by the Enactor engine of the ACGT Enactment Environment [18]. This inherent parallelization along with the usage of Grid resources, both in computation and storage terms, permits simultaneous execution of the same workflow with different parameters or input data resulting in significant up scale of the execution speed and the overall performance.

The implemented scenario was applied on an indicative Wilm's tumor gene-expression study [21]. We target the apoptosis GRN (KEGG identifier: hsa04210) as it engages prominent regulatory mechanisms for various cancer types and tumor states. The resulted significant paths are shown in Figure 2. As it can be observed, all the paths have a tendency to move towards the "Survival" or the "Death Genes" regions of the pathway. A pathological situation that leads constantly to the "Survival" or the blocking of the "Death Genes" region can be presumed to lead into endless proliferation of the cell. This finding can act as an indication for further analysis and research in this area.

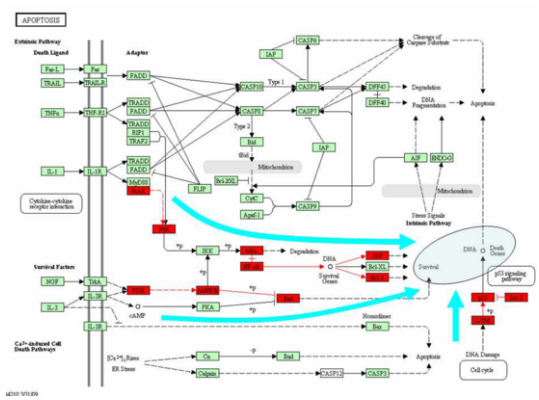

Figure 2 - If we portray the identified paths (red frames and connectors) as a layer on top of the apoptosis GRN then three primaries generic paths (light blue arrows) are emerged that lead towards cellular segments that are known to play vital role in the development (or suppression) of various types of cancer.

Discussion

The design and implementation of the workflow using the ACGT Enactment Environment, enables us to meet most of the above mentioned functional requirements

- The execution of the workflow is performed in a secure manner, using a Public Key Infrastructure (PKI) and access to private data is done only from authorized users. All activities are accessed using the end user Grid credentials and therefore authentication, authorization, and privacy are guaranteed.
- The design of the workflow is intuitive, following a data-flow analysis of the scenario and designing it using a "drag 'n' drop" web based visual editor. In particular the use of web technologies enables ubiquitous access and the straightforward sharing of the constructed scientific workflows.
- The decomposition of the GRN scenario to smaller functional steps enables the composition of the workflow by re-using existing ACGT tools and services, invoking also third-party services and accessing publicly available data, along with the private data on which only the user has access.
- The use of metadata, service, and domain ontologies, although not explicitly demonstrated in this scenario, enables the high level composition and orchestration of data analysis tasks.
- And finally the usage of grid resources and the parallel execution of various BPEL tasks, both of which has a proven record in terms of performance and scalability, results in a reliable execution of the scenario, which scales in a logical manner along with the scale of the input data.

Conclusion

The presented scenario tries to combine two already known sources of biomedical information in order to produce novel knowledge. The decomposition can be perceived as a disassembling of a partly known device into chunks of sub-mechanisms that can be more easily be tested and verified. Our testing platform is the microarray experiments conducted for certain disease types. In our application to Wilm's tumor, the identified paths revealed a more general and abstract path that might give more insights in the genomic regulations that happen during the advancement of the disease. The methodology does not only present an approach for the functional enrichment of microarray data - based on their abundance in specific pathways but also, it is able to reveal and identify regulatory-mechanisms (paths in gene regulatory networks) that discriminate and 'govern' the expression of specific phenotypes.

The presented methodology was also applied and tested on various clinico-genomic studies. In [22] it is utilized in the context of the breast-cancer (BRCA) domain and the gene expression profiling of BRCA patients targeting the Estrogen Receptor (ER) phenotypic categories. In [23] the methodology was applied on a well-known microarray study that targets the distinction between AML (Acute Myeloid Leukemia) and ALL (Acute Lymphoblastic Leukemia) leukemia sub-types. The ACGT scientific workflow environment was also utilized in the context of a real-world genotyping study that targets the discrimination between BRCA and normal patient samples through the identification of 100 SNPs (single nucleotide polymorphisms), measured with the Affymetrix 10K platform [24].

In this exercise the ACGT technologies and tools have succeeded in implementing such a scenario to its entirety and this fact gives more validity to the specific choices we have made and the approach we have followed. Nevertheless we acknowledge the fact that future work is needed in order to enhance the functionality and the utility of the system to the practicing bioinformatician. A particular aspect we are working on is with respect to the "reproducibility" of the workflow results which is strongly connected to the storage and management of the "provenance" information. Furthermore the use of graphical tools for the specification of complex experiments has its limitations. An experienced user may feel more "at home" if some scripting environment allows him/her to easily present his/her workflows in a declarative manner and it's on our agenda to research this direction further in the future.

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 in part by the European Commission (FP6/2004/IST-026996).

References

[1] Friend SH. How DNA microarrays and expression profiling will affect clinical practice. British Medical Journal. 1999; 319(7220):1306.

[2] Collins F, Lander E, Southern E. The chipping forecast. Nat Genet Suppl. 1999; 21:1–55.

[3] Belloum A, Deelman E, Zhao Z. Scientific workflows. Scientific Programming. 2006; 14(3-4):171.

[4] Zhao Z, Belloum A, Bubak M. Editorial: Special section on workflow systems and applications in e-Science. Future Gener Comput Syst. 2009; 25(5):525–527.

[5] M. Tsiknakis, M. Brochhausen, J. Nabrzyski, J. Pucaski, S. Sfakianakis, G. Potamias, C. Desmedt and D. Kafetzopoulos, A semantic grid infrastructure enabling integrated access and analysis of multilevel biomedical data in support of post-genomic clinical trials on Cancer, IEEE Transactions on Information Technology in Biomedicine, 2008, vol 12, no 2, pp. 191-204.

[6] Kanterakis A, Moustakis V, Kafetzopoulos D, Potamias G. Revealing Disease Mechanisms via Coupling Molecular Pathways Scaffolds and Microarrays: A Study on the Wilms Tumor Disease. In: Bassiliades N, editor. Workshops of the 5th IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI-2009). vol. 475 of CEUR Workshop Proceedings. CEUR-WS.org; 2009. p. 88–99.

[7] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000; 28(1):27.

[8] Kauffman S. The large scale structure and dynamics of gene control circuits: an ensemble approach. Journal of Theoretical Biology. 1974; 44(1):167.

[9] Pensa RG, Leschi C, Besson J, Boulicaut JF. Assessment of discretization techniques for relevant pattern discovery from gene expression data. In: Proceedings ACM BIOKDD; 2004. p. 24–30.

[10] Potamias G, Koumakis L, Moustakis V. Gene Selection via Discretized Gene-Expression Profiles and Greedy Feature-Elimination. In: SETN. vol. 3025 of Lecture Notes in Computer Science. Springer; 2004. p. 256–266.

[11] Foster I. The grid: computing without bounds. Scientific American. 2003; 288(4):78–85.

[12] Oinn T, Greenwood M, Addis M, Alpdemir MN, Ferris J, Glover K, et al. Taverna: Lessons in creating a workflow environment for the life sciences. Concurrency and Computation: Practice and Experience. 2006; 18(10).

[13] Altintas I, Berkley C, Jaeger E, Jones M, Ludäscher B, Mock S. Kepler: An extensible system for design and execution of scientific workflows. In: Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on; 2004. p. 423–424.

[14] Taylor I, Wang I, Shields M, Majithia S. Distributed computing with Triana on the Grid. Concurrency and Computation: Practice and Experience. 2005; 17(9).

[15] Gil Y, Deelman E, Ellisman M, Fahringer T, Fox G, Gannon D, Goble C, Livny M, Moreau L, Myers J. Examining the challenges of scientific workflows. Computer. 2007; 40(12):24–32.

[16] Pukacki J, Kosiedowski M, Mikolajczak R, Adamski M, Grabowski P, Jankowski M, et al. Programming grid applications with Gridge. Computational Methods in Science and Technology. 2006; 12(1):47–68.

[17] Kashyap V, Sheth A. Semantic heterogeneity in global information systems: The role of metadata, context and ontologies. Cooperative Information Systems: Current Trends and Directions. 1996; p. 139–178.

[18] Sfakianakis S, Koumakis L, Zacharioudakis G, Tsiknakis M. Web-Based Authoring and Secure Enactment of Bioinformatics Workflows. Grid and Pervasive Computing Conference, Workshops at the. 2009; 0:88-95.

[19] Alves A, Arkin A, Askary S, Barreto C, Bloch B, Curbera F, et al. Web services business process execution language version 2.0. OASIS Standard. 2007; 11.

[20] Wilkinson MD, Links M. BioMOBY: an open source biological web services proposal. Briefings in bioinformatics. 2002; 3(4):331–341.

[21] Zirn B, Hartmann O, Samans B, Krause M, Wittmann S, Mertens F, et al. Expression profiling of Wilms tumors reveals new candidate genes for different clinical parameters. International Journal of Cancer. 2006; 118(8).

[22] Kanterakis, A., Kafetzopoulos, D., Moustakis, D., and Potamias, G. Mining Gene Expression Profiles and Gene Regulatory Networks: Identification of Phenotype-Specific Molecular Mechanisms. Lecture Notes in Artificial Intelligence, 5138: 97-109.

[23] Kanterakis, A., Kafetzopoulos, D., Moustakis, V., and Potamias, G. Mining Gene Regulatory Networks and Microarray Data: The MinePath Approach. 18th European Conference on Artificial Intelligence (ECAI'08), BIGPAIA workshop, July 22, 2008, Patras, Greece

[24] Koumakis, L., Sfakianakis, S., Moustakis, V., and Potamias, G. Discovery of Genotype–to–Phenotype Associations: A Grid–enabled Scientific Workflow Setting. BMIINT: Biomedical Informatics & Intelligent Methods in the Support of Genomic Medicine, AIAI 2009 workshop, April 24 2009, Thessaloniki, Greece, CEUR Proceedings 475:48-59

Address for correspondence

Dr. George Potamias, Biomedical Informatics Laboratory, FORTH-ICS, Vassilika Vouton P.O Box 1385, GR-71110 Heraklion, Crete, Greece, Phone: +30 2810 391693, Fax: +30 2810 391428,
E-mail: potamias@ics.forth.gr