

Inferencing in *In Silico* Oncology: Exploiting Expressions, Biomarkers and Clinical Data for Clinical Decision Support

D. I. Fotiadis, Y. Goletsis, T.P. Exarchos, N. Giannakeas and G. Rigas

Abstract—The large number of bioinformatics applications during the recent years offers an abundance of experimental data related to oncology. This leads to the need for efficient algorithms and computational techniques that integrate different types of data (coming from different sources) and derive knowledge out of a the evolving volume of data. In this paper we demonstrate some of our recent research work towards inferencing in *In Silico* Oncology.

Keywords: gene expressions, genetic networks, genetic sequence pattern analysis, data mining,

Bioinformatics, i.e. the creation and advancement of algorithms, computational and statistical techniques, and theory to solve formal and practical problems posed by or inspired from the management and analysis of biological data, during the last years offers a large number of approaches applicable to oncology. However, due to the abundance of experimental data, efficient algorithms and techniques which integrate different types of data (coming from different sources) and derive knowledge out of a significant and continuously evolving volume of data are needed in order to support clinicians/biologists in their medical practice. Inferencing out of these data include the exploitation of gene expressions, demographic, clinical data including biomarkers as well as sequence data.

Processing of microarray images typically usually consists of gridding and spot finding, segmentation and intensity extraction [1,2]. Several difficulties appear in all above steps, such as variations of block and spot positions, existence of non-expressed spots that have zero intensity, existence of dust or other contamination on the slide that generates artefacts in the image. For this reason efficient computational analysis tools and techniques are required.

In [3] we introduced a novel five-step method to deal with all these difficulties. Initially, the raw microarray image is preprocessed with a template matching technique. In the second step, the blocks of the image are located. The third step is the spot finding in each block. In this step outlier detection is applied on each row and column of spots

in order to remove the artefacts. The next step is the detection of the non-expressed spots. Finally a grid is fit on the image using a Voronoi diagram. The results of the above method are promising and this will lead us in the development of an effective segmentation technique in order to accurately extract the gene's expression.

Gene Expression data can be exploited for the reconstruction of the genetic networks (see Fig.1). Probabilistic methods can be used for inferring complex relations between genes. In this way, regulatory mechanisms will be inferred and protein functions can be revealed. Bayesian networks have been long examined as a prominent approach for deriving network structure [4,5]. Still, there are a number of challenges that remain open. The first challenge is how to handle missing values in the procedure of learning Bayesian networks. For this reason, in one of our recent works, we employed the Structural EM [6], to handle missing values in the learning of the Bayesian network's structure. The Bayesian network recovers the structure of regulatory interactions between the different genes.

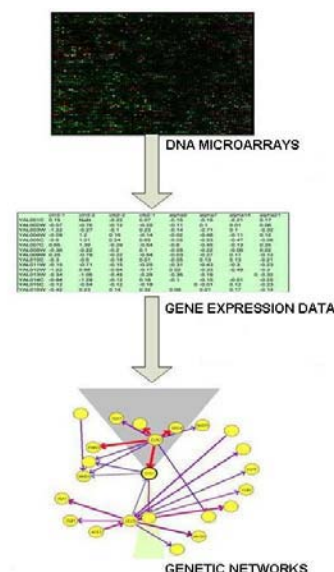


Fig. 1: Reconstruction of genetic networks from microarray data.

A major drawback of the above procedure is the need for high computational effort, even for relatively small networks. The solution we proposed is a parallel implementation of the whole procedure. Following several tests, our results are quite satisfactory demonstrating the efficiency of our approach in genetic network reconstruction. Our results also demonstrate the need for incorporation of GRID technologies in expression data analysis.

Besides gene expression data, sequence data could be useful for *In Silico* Oncology applications. An innovative

D. I. Fotiadis⁺, T.P. Exarchos and G. Rigas are with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, Ioannina, Greece, GR 45110

Y. Goletsis is with the Dept. of Economics and with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, Ioannina, Greece, GR 45110

N. Giannakeas is with the Laboratory of Biological Chemistry, Medical School and the Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science University of Ioannina, Ioannina, Greece, GR 45110.

⁺ corresponding author (e-mail: fotiadis@cs.uoi.gr).

project, MATCH [7], investigates how data mining and pattern recognition techniques can combine clinical and sequence data for profiling and diagnostic purposes in colon cancer. MATCH is developing a web based multi functional platform that integrates medicine and molecular biology to provide more effective treatment and enhance pharmaceutical research and drug discovery. In MATCH, clinical and biological data are integrated in order to (i) discover correlations between SNPs and colon cancer and (ii) allow for patient diagnosis, staging and treatment selection.

Data integration is the key to MATCH platform. Clinical data derived from the electronic healthcare records and biological information derived using data mining techniques from patient genomic and proteomic sources, are analyzed in order to provide patient profiles. A specially built Decision Support System matches new patient data (demographic, clinical, genetic) against these profiles. The relationship between sequence profiles under different experimental conditions and biological processes can be drawn through pattern analysis. This newly designed data mining model provides an efficient way to translate the large collection of existing profiles so as to be a handy reference for clinicians who face cancer early detection, clinical diagnosis and treatment decisions.

In MATCH architecture, a special ontology undertakes the role of facilitating knowledge sharing and reuse, while it will be experimentally integrated in the decision making process, the latter being a quite innovative feature [8]. MATCH ontology is built on top of already existing and proved ontologies (i.e. GeneOntology [9], National Cancer Institute ontology [10], Sequence Ontology [11]).

The architecture of the MATCH project is shown in Figures 2 and 3.

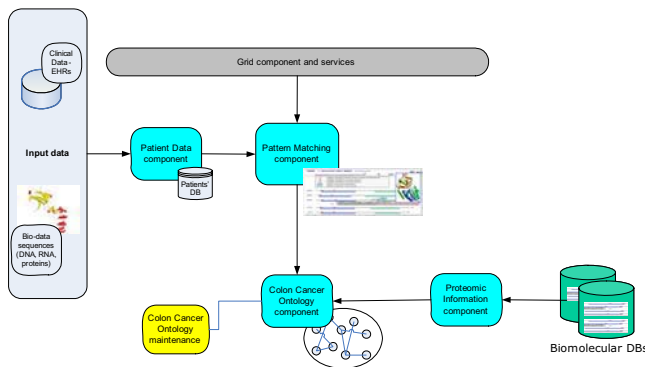


Fig. 2: MATCH architecture – components participating in training phase.

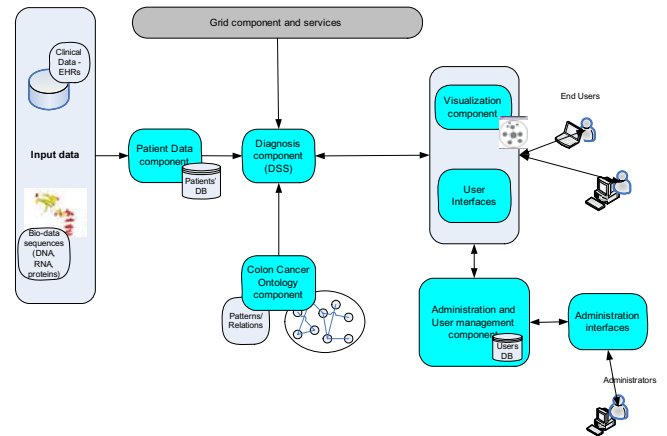


Fig. 3: MATCH architecture – components participating in decision support phase.

REFERENCES

- [1] Yang, Y.H. Buckley, M.I. Dudoit, S. and Speed, T.P. Comparison of methods for image analysis on cDNA microarray data. *J. Computational and Graphical Stat.*, 11, 108-136, 2002.
- [2] Jain, A.N. Tokuyasu, T.A. Snijders, A. M. Segraves, R. Albertson, D.G. and Pinkel, D. Fully automatic quantification of microarray image data. *Genome Res.*, 12, 325-332, 2002.
- [3] Giannakeas, N. Fotiadis, D.I. and Politou, A.S. An Automated Method for Gridding in Microarray Images, proceedings of 28th IEEE EMBS conference, 5876-5879, 2006.
- [4] Friedman, N. Linial, M. Nachman, I. Pe'er, D. Using Bayesian networks to analyze expression data. *RECOMB 2000*, 127-135, USA, 2000.
- [5] Spirtes, P. Glymour, C. and Scheines, R. Constructing Bayesian network models of gene expression networks from microarray data, 2000.
- [6] Friedman, N. Learning belief networks in the presence of missing values and hidden variables. *Proc. 14th Int. Conf. Machine Learning*, 125-133, 1997.
- [7] MATCH: Automated diagnosis system for the treatment of colon cancer by discovering mutations on tumor suppressor genes, project IST-2005-027266, website <http://www.match-project.com/>
- [8] Ceccaroni, L. Cortes, U. Sanchez-Marre, M. OntoWEDSS: augmenting environmental decision-support with ontologies, *Environmental Modelling & Software*. 19, 785-797, 2004.
- [9] <http://www.geneontology.org/>
- [10] <http://www.mindswap.org/2003/CancerOntology/>
- [11] <http://song.sourceforge.net/>