34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Multi-scale modeling of Gene Regulatory Networks via integration of temporal and topological biological data *

George Dimitrakopoulos, *Student Member, IEEE*, Kyriakos Sgarbas, *Member, IEEE,* Konstantina Dimitrakopoulou, *Student Member, IEEE*, Andrei Dragomir, Anastasios Bezerianos, *Senior Member, IEEE* and Ioannis A. Maraziotis

Abstract— **Regulome is the dynamic network representation of the regulatory interplay among genes, proteins and other cellular components that control cellular processes. Reconstruction of gene regulatory networks (GRN) delineates one of the main objectives of Systems Biology towards understanding the organization of regulome. Significant progress has been reported the last years regarding GRN reconstruction methods, but the majority of them either consider information originating solely from gene expression data or/and are applied on a small fraction of the experimental dataset. In this paper, we will describe an integrative method, utilizing both temporal information arriving from time-series gene expression profiles, as well as topological properties of protein networks. The proposed methodology detects relations among either groups of genes or specific genes depending on the level of abstraction or resolution requested. Application on real data proved the ability of the method to extract relations in accordance with current biological knowledge as well as discriminate between different experimental conditions.**

I. Introduction

Modeling and simulation of the regulome delineates one of the main objectives of Systems Biology towards understanding the functional organization of cells and the mechanisms by which mis-regulation leads to certain diseases [1,2].

Advances in molecular high-throughput techniques have generated vast amounts of data for genes and proteins that are the two key factors in regulome. Hybridization methods like DNA microarrays, provide gene expression measurements that offer the ability to monitor the behavior for thousands of genes under many experimental conditions. Protein-protein interaction (PPI) data descending from various techniques depict the physical interactions governing cellular processes.

Bioinformatics studies over the last decade provided algorithmic methodologies that attempted to attribute different manifestations of the regulome focusing on one kind of molecular interactions (i.e. gene-gene, protein-protein). Specifically, most of the relevant literature regarding regulome modeling is divided into two basic directions, namely Gene Regulatory Networks (GRN) reconstruction dealing with the interrelations among genes and functional modules detection that extracts sets of closely functionally related proteins.

Initial approaches on the GRN reconstruction problem exploited gene expression data arriving mainly from microarray experiments [3,4]. However, challenges like the under-determinism caused by the large number of variables and the complexity introduced by the expanded search space (i.e. groups of genes co-operate to activate or repress a target gene), prohibit data-driven computational models to predict accurately large scale GRNs and restrict their application to only a few tens of genes.

Later studies attempted to overcome these constraints by applying various compression techniques [5] that, while reduced the computational complexity, led to valuable information loss. In other cases, various heuristic schemes concerning the selection of possible gene regulators were integrated as pre-processing steps in the main algorithms of reconstruction [6]. While approaches such as these contributed towards more efficient both in time and accuracy final solutions, still the limiting factor that they act solely upon gene expression data, prevents them from revealing the actual biological "portrait" [7].

The cornerstone of the proposed methodology is the integration of temporal gene expression and PPI data, in the form of a composite weighted graph, to increase the robustness of the derived results. An important attribute of the reformed graph is the representation of casual interactions among individual genes or groups of genes based on the requested level of abstraction. Additionally, we present a framework to cope with the high dimensionality of microarray experiments for retaining all gene expression information, thus enabling a realistic large scale overview of the transcriptome network.

We applied our approach in a known therapeutic agent, interferon-beta (IFN-β), in controlling exacerbations in relapsing - remitting multiple sclerosis [8] with the scope to elucidate its mechanism of action. Little is known, regarding the mechanism by which IFN-β exerts its immunomodulatory effect at the cellular and organismal level. Our findings report a potential mode of action of this agent and reveal hidden pathway interconnections that warrant further study.

* This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thalis. Investing in knowledge society through the European Social Fund.

G.D. and K.S. are with the Department of Electrical and Computer Engineering, University of Patras, Patras, 26500, GR (geodimitrak@upatras.gr , sgarbas@upatras.gr).

I.A.M., K.D. and A.B. are with the Medical School, University of Patras, Patras, 26500, GR (imaraziotis@gmail.com, kondim@upatras.gr, correspondence author: +302610969145, bezer@upatras.gr).

A.D. is with the Department of Biomedical Engineering, University of Houston, Houston, TX 77204-5060, USA (adragomir@uh.edu)

II. METHODS

As we already mentioned, the first stage of the proposed methodology constructs a composite weighted graph based on PPI and gene expression data. Specifically, the nodes of the graph (henceforth called cnodes) represent groups of genes with similar expression profiles and originate from a standard clustering process. An edge between a pair of cnodes and its corresponding weight delineates the existence and the exact number of protein interactions shared among all members of the implicated cnodes. At this point, we have to mention that, in contrast to other methods that use clustering as a lossy compression process to reduce data dimensionality [9], we have applied clustering only as a form of data grouping, while we retain and employ all gene expression profiles present in the dataset. The large clustering resolution we have imposed in the original gene expression dataset and hence the large similarity among the members in every cluster, is biologically justified by the assumption that similar gene expression trends imply in many cases functional association in the protein level. Furthermore, a large number of PPIs between a pair of cnodes (expressed by a high corresponding weight value) indicates an analogous degree of biological similarity not only between the cnodes forming the pair, but also among the members of each one. Under a supervised learning point of view, the highly similar gene expression profiles in every cluster allow us to consider them as noisy manifestations of the same entity described by the centroid of the cluster. Hence, members of a certain group can be employed to formulate training and test datasets of the same entity (centroid) in a machine learning method.

The last assumptions allow us to provide a pair of interconnected cnodes as input to a method for GRN reconstruction based on an evolutionary trained multi-layer neuro-fuzzy neural network (ENFRN). The method described in detail in [6], accepts as input a set of gene expression profiles as training and test datasets and automatically determines the best potential regulators and target genes.

The ENFRN-based methodology can detect various types of simple or complex relations among genes based on their expression levels. This property originates from the capacity of the network to describe input / output relations based on a set of fuzzy rules. In this study, for reasons of simplicity the adapted types of relations are: activation (A->B or B->A), inhibition (A-|B or B-|A) or unspecified (A-B), where A and B represent a single gene or a small gene set along with a score indicating the confidence of the detected relation. Interactions bearing a score above a certain threshold are retained while the rest of them are disregarded. Finally, the type of relation between two cnodes is based on a maximum vote scheme concerning the corresponding types of the valid regulations extracted among the members of the cnodes.

A. Data

We used a time series microarray dataset studying the longitudinal transcriptional profile of blood cells within a week of IFN-β administration to multiple sclerosis patients, accessible through NCBI's Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) with series accession number GSE5678 [8]. The original dataset contains over 22000 gene probes whose levels of expression were measured across 6 different individuals, two patients treated with IFN-β (T1 and T2), and four untreated healthy (control) persons (U1-U4). The blood samples from T1, U1 and U3 were obtained at baseline and at 3.5, 6.25, 9.5, 11.5, 16.5, 25, 49 and 156 h and from T2, U2, U4 at 7.25, 10.25, 13, 15.5 and 33 h.

TABLE I.

Stages	Patient		Control	
	Nodes	*Edges*	*Nodes*	*Edges*
Initial dataset	9111	41928	9111	41928
Weighted graph	161	10734	159	10696
Thresholded weighted graph	125	437	114	495
Reconstructed directed graph	125	242	114	140
Genes of interest	27	67	16	25

The table contains information regarding the various operational stages of our proposed methodology and its progressive ability to discriminate between the 2 different experimental conditions.

To construct a coherent and reliable PPI network, we combined PPI data from various sources: Ophid [10], MINT [11] and BioGRID [12] databases and [13], collecting in total 59403 interactions among 11266 proteins.

After removing all genes that did not show significant level of differential expression across the two experimental conditions, we ended up with 11085 unique gene symbols or names. Lastly, we kept only the genes and proteins with common names, in both the gene expression and PPI datasets, thus resulting to a final number of 9111 genes/proteins, with 41928 corresponding PPIs (as shown in Table 1). The original gene expression dataset was split into two datasets, each one consisting of 14 time points, for the patient (combining T1 and T2 data) and the control (combining the average of U1 and U3 with the average of U2 and U4) case respectively.

B. Weighted Graph

In accordance to the aforementioned concepts, we initiated the construction of the two composite weighted graphs corresponding to the patient and control conditions, by partitioning each gene dataset with k-means algorithm into a large number of clusters. At this point, we stress out the dual goal fulfilled by clustering procedure. On one hand, we demand clusters with high similarity degree, while on the other each cluster must include a sufficient number of members to serve the needs of the forthcoming reconstruction method for training and test datasets. Hence, the cluster number was set initially to 150 for each dataset and oversized clusters were recursively split with threshold of 100 members each. This procedure resulted in 161 clusters for the patient dataset and 159 clusters for the control.

Next, we exploited the PPI dataset, in order to detect the interactions among cnodes. In detail, two weighted graphs were constructed by defining the weight of an edge as the number of PPIs between two cnodes. The two resulting networks contained 10734 edges for the patient dataset and 10696 edges for the control dataset with the maximum value of the weight being 127. To select biologically interesting cases, we considered a pair of cnodes connected, only if they shared more than 10 PPI relations. The final outcome was a graph with 437 edges connecting 125 cnodes for the patient

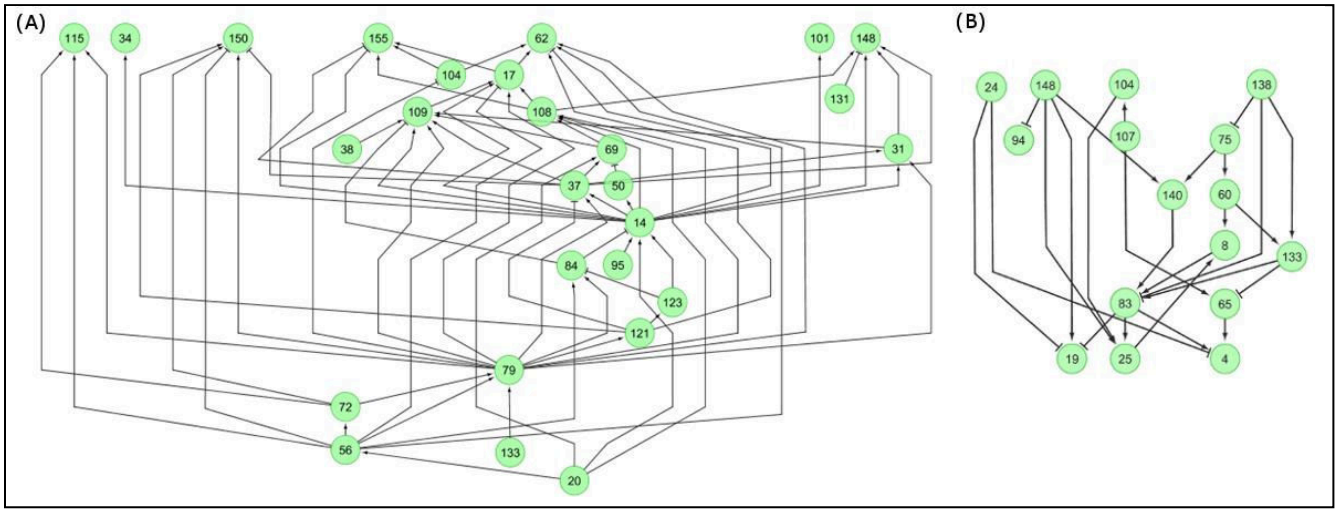

Figure 1. Subnetworks of patient (A) and control (B) states depicting interactions among a fraction of cnodes (cluster nodes) related to IFN-β mechanism. In (A) 27 cnodes (out of the total of 161) are present and in (B) 16 cnodes (out of 159) respectively. It is evident that the control graph displays low interactivity in comparison to the patient, with the latter condition indicative of the triggering of the therapeutic agent mode of action. (The cnodes are enumerated based on the clustering procedure of the two datasets.)

case and 495 edges connecting 114 cnodes for the control.

For each one of the two different experimental conditions, we defined as training set the 50% of farthest genes from the centroid of each cnode based on their Euclidean distance and as testing set the other 50% of the nearest ones. These values were selected based on the operation of the ENFRN methodology [6] and to avoid over-training problems. These sub-datasets were provided as input to the ENFRN-based reconstruction method as previously described.

Regarding the patient graph, out of the 437 edges 179 are characterized as up-regulation, 63 as down-regulation and 195 as undirected relations and regarding the control graph out of 495 edges as 102, 38, 355 respectively. We observe that 44.6% and 71.7% of the edges of each case are undirected. An undirected relation implies that the corresponding cnodes profiles show similar expression trends, thus they cannot be classified as up or down regulation.

III. Results and Discussion

As mentioned earlier, the proposed methodology can provide information of the complete transcriptome graph via interactions among cnodes as well as zoom in specific areas of interest concerning individual genes/proteins through the interactions among cnodes. To test the modeling efficiency of our methodology, we focused on a specific subset of the two reconstructed graphs (patient and control). Specifically, based on information gathered from relative literature [8,14,15], we compiled a list consisting of 79 genes implicated in both IFN-β mechanism and multiple-sclerosis disease. The selected genes are enriched with Gene Ontology (GO) terms such as immune response, apoptosis, protein biosynthesis and others, which are consistent with the canonical path described for IFN-β signaling and function. A second list was obtained from TFCat database [16] which included 252 genes that belonged to our dataset and are known to be transcription factors. In both cases, we constructed a subnetwork by selecting the cnodes that included at least one gene that either belonged in the first list or was transcription factor in a hub cnode. A cnode is characterized as hub if it has with N_h or more directed edges. In this study, we selected N_h equal to 10 (trials in the range between 5 and 10 gave the same results) and we concluded in a subset of cnodes with 8 and 7 genes for patient and control dataset respectively.

The resulting patient network was densely connected and consisted of 27 cnodes with 67 edges, while the corresponding control network consisted of 16 cnodes with 25 edges (Fig. 1). We observed in the patient graph enhanced interactivity among cnodes that mainly implicated genes and pathways related to the IFN-β regulatory effect on transcriptome. On the contrary, the majority of the extracted relations among different cnodes in the control reversed engineered graph mainly reflected cellular physiological processes, while its hub cnodes mainly contained housekeeping genes.

Moving forward, we further analyzed the cnode graphs by classifying the cnodes in GO categories based on the GO terms of the genes of interest that they contain. In the control case, we observed that the majority of genes (9 out of 16) belonged to the GO term transcription from RNA polymerase II promoter, while the rest to other general physiological process terms, thus the graph was degenerated in a list of 3 nodes (data is not shown). On the other hand, the resulting patient graph (Fig. 2) displayed enhanced interconnectivity among the GO categories, indicating so that a strong external perturbation (an immunomodulatory drug) results in an increased and coordinated recruitment of several pathways. Additionally, we corroborate that the terms apoptosis and transcription from RNA polymerase II promoter have a key role in IFN-β mechanism which is consistent with biological observations that many processes converge into programmed cell death after IFN-β administration.

In Fig. 3, we zoom into three GO categories of the aforementioned graph and provide characteristic examples of gene-gene interactions that we successfully identified with respect to known KEGG pathways [17].

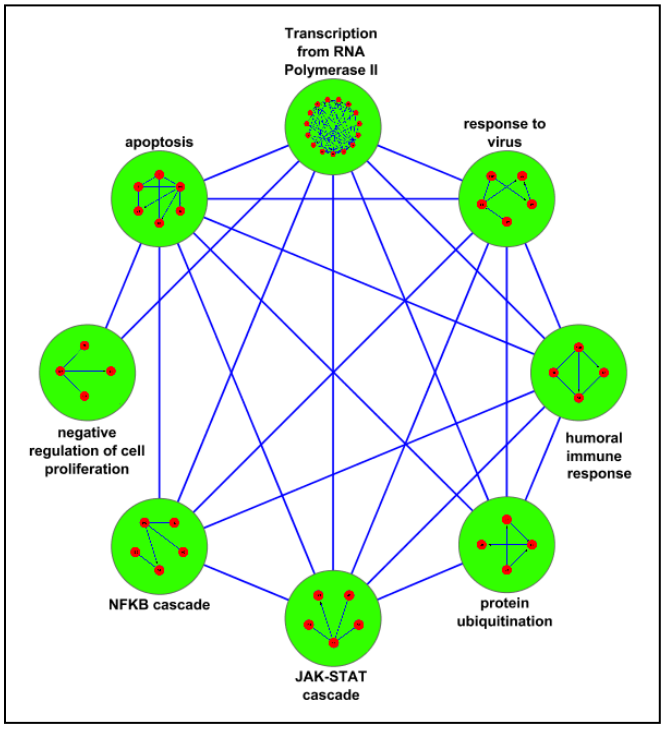

Figure 2. GO-based graph depicting the patient state. The red nodes in the nested networks represent cnodes containing IFN-β related genes. The cnodes sharing the same GO term are grouped in one large green node. The high density of the graph reflects the enhanced interactivity among pathways.

IV. Conclusion

Initial experiments proved that the proposed methodology is an efficient method able to provide results in accordance with biological knowledge. In contrast to other methods that can only be applied in a small fraction of genes contained in a microarray experiment, our method can operate under full scale experimental conditions. Its ability to provide information under different levels of abstraction can be a valuable resource towards deciphering and better understanding the complex modular relations among genes and proteins.

In future work, we plan to incorporate functional module extraction algorithms such as DMSP [18], able to operate on individual 'seed' cnodes to extract specific subnetworks, and hence further enhance the ability of the method to focus on pathways of interest depending on the biological experiment. Furthermore, one of the most important steps of the proposed methodology is the clustering of gene expression data. As the degree of clustering resolution (i.e. number of clusters) is increased the noise level in every cluster is decreased and hence the biological signal is better preserved. However, algorithms like k-means, which was employed in this study, fail to adequately operate under large degrees of resolution. We are working on algorithms that can address this problem and hence further enhance the framework we proposed.

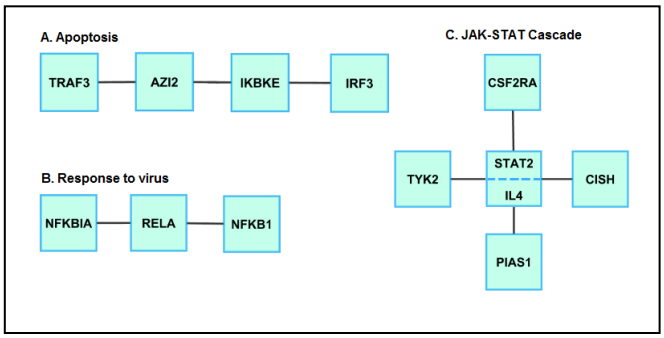

Figure 3. Characteristic examples of correctly identified gene/protein relations belonging to three GO terms (extracted from Fig. 2) associated with IFN-β signaling pathway. The case of STAT2 and IL4 that are located in the same box, indicates that those two proteins are members of the same cnode.

References

[1] O. X. Cordero, and P. Hogeweg, "Large changes in regulome size herald the main prokaryotic lineages", *Trends in genetics*, vol. 23, pp. 488-493, Oct. 2007.

[2] G. K. Sandve, *et al.*, "The differential disease regulome", *BMC Genomics*, 12:353, Jul. 2011.

[3] W. P. Lee, and W. S. Tzou, "Computational methods for discovering gene networks from expression data", *Brief. Bioinform.*, vol. 10, pp. 408-23, Jul. 2010.

[4] I. A. Maraziotis, A. Dragomir and A. Bezerianos, "Gene networks reconstruction and time series prediction from microarray data using recurrent neural fuzzy networks", *IET Systems Biology*, vol.1, pp.41-50, Jan. 2007.

[5] R. Guthke, U. Moller, M. Hoffmann, F. Thies and S. Toepfer, "Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection", *Bioinformatics*, vol. 21, pp. 1626-1634, Dec. 2004.

[6] I. A. Maraziotis, A. Dragomir, and D. Thanos, "Gene Regulatory networks modeling using a dynamic evolutionary hybrid", *BMC Bioinformatics*, 11:140, Mar. 2010.

[7] R. Pal, S. Bhattacharya, and M. U. Caglar, "Robust Approaches for Genetic Regulatory Network Modeling and Intervention: A review of recent advances", *Signal Processing Magazine, IEEE*, vol.29, pp.66-76, Jan. 2012.

[8] G. H. Fernald, *et al.*, "Genome-wide network analysis reveals the global properties of IFN-beta immediate transcriptional effects in humans", *J. Immunol.*, vol. 178, pp. 5076-85, Apr. 2007.

[9] M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, and R. Guthke, "Gene regulatory network inference: data integration in dynamic models-a review", *Biosystems*, vol. 96, pp. 86-103, Apr. 2009.

[10] K. R. Brown, and I. Jurisica, "Online Predicted Human Interaction Database", *Bioinformatics*, vol. 21, pp. 2076-82, May 2005.

[11] A. Ceol, *et al.*, "MINT, the molecular interaction database: 2009 update", *Nucleic Acids Res.*, vol. 38, pp. 532-9, Jan. 2010.

[12] C. Stark, *et al.*, "BioGRID: a general repository for interaction datasets", *Nucleic Acids Res.*, vol. 34, pp. 535-9, Jan. 2006.

[13] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis", *Mol. Syst. Biol.*, 3:140, Oct. 2007.

[14] D. Miller, F. Barkhof, X. Montalban, A. Thompson, and M. Filippi, "Clinically isolated syndromes suggestive of multiple sclerosis, part I: natural history, pathogenesis, diagnosis, and prognosis", *Lancet Neurol.*, vol. 4, pp. 281-8, May 2005.

[15] R. A. Marrie, "Environmental risk factors in multiple sclerosis aetiology", *Lancet Neurol.*, vol. 3, pp. 709-18, Dec. 2004.

[16] D. L. Fulton, S. Sundararajan, G. Badis, T. R. Hughes, W. W. Wasserman, J. C. Roach, and R. Sladek, "TFCat: the curated catalog of mouse and human transcription factors", *Genome Biol.*, 10:R29, Mar. 2009.

[17] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets", *Nucleic Acids Res.*, vol. 40, pp. 109-114, Nov. 2011.

[18] I. A. Maraziotis, K. Dimitrakopoulou, and A. Bezerianos, "Growing functional modules from a seed protein via integration of protein interaction and gene expression data", BMC Bioinformatics, 8:408, October 2007.