

# Qualitative Modeling Framework for Identification of Genetic Regulatory Networks

*Rajanikanth Vadigepalli, Ronald K. Pearson, Daniel E. Zak, James S. Schwaber*

Many complex processes such as adaptation, development, differentiation, cell cycle etc. in biological systems involve multiple genes functioning in a hierarchical and highly interconnected structure. The advent of gene expression microarray technology in the recent years has resulted in high throughput data sets, the analysis of which holds the promise of identification of the nonlinear dynamic function of the biochemical regulatory networks. However, the precision and reliability of these quantitative data sets varies widely and most often permit only a coarse-grained analysis (up regulation, down regulation, no effect etc.). The conventional modeling approaches are not well suited for such coarsely quantized data sets and are likely to result in inconsistent, if not incorrect, network identification.

We have recently proposed a novel modeling methodology for genetic regulatory networks utilizing qualitative information alone [1]. This framework can deal with coarsely quantized and uncertain genomic and proteomic data that assumes values in the set  $\{+,-,0,*\}$  corresponding to {increase, decrease, no change, unknown}, respectively. This objective of this approach is to obtain qualitative information such as signs of interactions in the regulatory network;  $\{+,-,0,*\}$  corresponding to {activation, repression, no interaction, unknown}, respectively. The methodology is easily extensible to data sets that are quantized further, e.g.,  $\{++,+,0,-,--,*\}$ .

The model structure employed in identification is akin to the linear quantitative models, and is represented as:  $y=Zp$ , where  $Z$  contains the time series data of gene expression and transcription factor activity,  $y$  is the time series of predicted gene expression and  $p$  is the parameter vector of regulatory interactions. In this model class, all the data is quantized, i.e., elements of  $y$ ,  $Z$  and  $p$  are from the set  $\{+,-,0,*\}$ . The prediction error metric utilized in parameter identification is based on weighted dissimilarity between predicted and observed gene expression quantizations. The parameter identification is a nonlinear optimization problem with extremely large search space ( $N*4^N$  possible values for  $N$  genes) allowing for every gene to be regulated by all other genes. This is unrealistic for gene regulatory networks in biological systems.

We propose an approach where only a maximum of 'm' regulatory interactions are allowed thus drastically reducing the search space to  $N*({}^N C_m)*4^m$ , where  ${}^N C_m$  is the number of combinations of 'm' elements taken from 'N'. We also propose a systematic procedure based on integer-based optimization to render the parameter identification problem tractable.

The parameter sets for which the prediction dissimilarity measure is below a threshold are considered rather than a single 'optimal' solution. From these candidate sets of networks, we compute a 'consensus' network with confidence levels assigned to the regulatory interactions in the network. This approach is demonstrated on a simulated realistic genetic regulatory network

available in published literature [2], and is constructed from known biological regulatory structures.

1. R. Pearson, R. Vadigepalli, D.E. Zak and J.S. Schwaber. Qualitative Analysis in Systems Biology. *Submitted*.

2. D.E. Zak , F.J. Doyle, G.E. Gonye and J.S. Schwaber. Simulation studies for the identification of genetic networks from cDNA array and regulatory activity data. *Proceedings of the Second International Conference on Systems Biology*. 231-238, 2001.