
Stochastic Integrative Modeling of Transcription Regulation

Noa Novershtern

School of Computer Science and Engineering
Hebrew University
Jerusalem, Israel
shefi@cs.huji.ac.il

Aviv Regev

Broad Institute of MIT and Harvard
7 Cambridge Center, Cambridge, MA 02142
aregev@broad.mit.edu

Nir Friedman

School of Computer Science and Engineering
Hebrew University
Jerusalem, Israel
nir@cs.huji.ac.il

Keywords: Transcription, Module, Network, Integrative, Interactions

1 Introduction

Modeling the complex mechanisms of transcription regulation is a major challenge in computational biology. Different approaches have been used to address this challenge, one of which is the probabilistic approach, often represented by *Bayesian Networks* that explicitly describe the uncertainty which is inherent to biological data (Friedman et al. [2000], Hartemink et al. [2001]). Later extensions of Bayesian Networks, such as *Module Networks* (Segal et al. [2003a, 2005]), use the modular organization of transcriptional responses to reconstruct modules of co-expressed genes and their “regulation programs” — a set of regulators that control the modules’ activity.

These sophisticated models are nevertheless limited in using gene expression as a sole source of observations. The increasing availability of large-scale measurements of other molecular data types, such as protein-protein interactions and sequence data, facilitates the reconstruction of integrative models. The integration of multiple data types not only results in a richer model that describes several dimensions of the biological picture, but can also yield a more accurate model that bases its hypotheses on several independent sources of data.

Indeed, several recent studies have used different types of data to reconstruct such models (among them Ideker et al. [2002], Segal et al. [2002, 2003b], Yeang et al. [2004], Nariai et al. [2005]). However, most of these models focus on one dimension which is stochastic, *i.e.* contains free parameters, while the other types of information are used merely as constraints.

Here we present *Physical Module Networks*, a model that integrates protein interactions, protein-DNA binding data and gene expression data into a unified probabilistic scheme. Unlike most existing approaches, our model is stochastic in all dimensions, and enables us to infer the transcription network as well as the physical network underlying the biological observations.

2 Physical Module Network

The Physical Module Network (PMN) model consists of two components, one represents transcription modules and their regulation programs, and the other represents protein-protein interactions and protein-DNA binding.

The first component of the model is a Module Network $\mathcal{M}n$ (Figure 1 (a)). This is a Bayesian network over a set of modules $\mathcal{C} = \{M_1 \dots M_k\}$. Each module represents a group of genes that share a set of parents $Pa_{M_j} \in \mathcal{C}$, and whose expression levels are derived from the same distribution, given the value of their parents. Therefore, the genes in the module share their conditional probability distribution (CPD) $P(M_j | Pa_{M_j})$.

The second component of the PMN is a *Physical Interaction Map* $\mathcal{P}h$ (Figure 1 (b)). This is a partially directed graph over a set of genes and their product proteins that describes the physical interactions between proteins and genes. It consists of three types of edges: An undirected protein-protein interaction edge, a directed protein-DNA binding edge and a directed production edge from a gene to its protein product. The genes in the interaction map do not necessarily appear in the module network.

While the module network describes how a regulator affects the transcription of its target genes, the physical interaction map describes the set of interactions that may lead to this effect. Thus, to present a coherent picture of the full regulatory mechanism we must ensure that these two components are consistent. This is achieved by a set of rules that define which interactions are “legal” given the modules’ configuration. Specifically, we require that a regulator of a module will have a path of physical interactions to all the genes in that module. This *Regulation Path*, which begins with the protein node of the regulator, should end with a transcription factor that binds all the genes in the module (Figure 1).

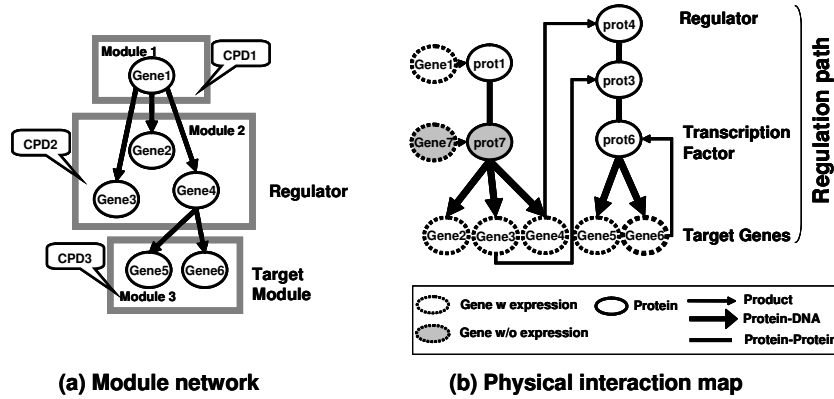


Figure 1: A simple scheme of a PMN model. This model integrates two consistent components: a Module Network (a) and a Physical Interaction Map (b). Note that both regulators Gene1 and Gene4 have a legal interaction path to the target genes in the interaction map.

3 Reconstructing the Network

To find the best model given the data, we use a greedy search heuristic in which we make local changes in the modules’ composition and in their choice of regulators, where each change is made simultaneously in the module network and in the interaction map. The search target function is defined as the Bayesian score, which is derived from the posterior probability of the model. Under certain assumptions the score decomposes into the two components:

$$\begin{aligned}
 \text{Score}(\mathcal{M}p|D) &= \log P(\langle \mathcal{M}n, \mathcal{P}h \rangle | \langle \mathcal{X}, \mathcal{I} \rangle) + c \\
 &= \log P(\langle \mathcal{M}n, \mathcal{P}h \rangle) + \log P(\langle \mathcal{X}, \mathcal{I} \rangle | \langle \mathcal{M}n, \mathcal{P}h \rangle) + c \\
 &= \log P(\mathcal{M}n) + \log P(\mathcal{P}h) + \log P(\mathcal{X} | \mathcal{M}n) + \log P(\mathcal{I} | \mathcal{P}h) + c
 \end{aligned}$$

Where $\log P(\mathcal{M}n)$ and $\log P(\mathcal{P}h)$ are the prior beliefs in the module network and the interaction map, respectively; $\log P(\mathcal{X} | \mathcal{M}n)$ is the likelihood of the gene expression observations \mathcal{X} ; and $\log P(\mathcal{I} | \mathcal{P}h)$ is the likelihood of the physical interaction observations \mathcal{I} .

The interaction observations are confidence levels assigned to each interaction measurement, where interactions that are not supported by the biological measurements are also allowed in our map, assigned with a low default probability. This enables, for instance, to connect each transcription factor to all the genes in a certain module.

To ensure the model's consistency throughout the search, we define the prior for an inconsistent model to be zero. In practice, we avoid taking steps in the search space that violate the consistency. For instance, when choosing a regulator r to regulate a module M , *i.e.* adding an edge in the module network, we need to find a pathway of interactions from r to the genes in M that also maximizes the Bayesian score. We reconstruct such a pathway with weighting each edge in the interaction graph with the change in the score upon the addition of that edge to the model $W(e) = \Delta \text{Score}(\mathcal{P}h|\mathcal{I}) = \log P(e \in \mathcal{P}h|\mathcal{I}(e)) - \log P(e \notin \mathcal{P}h|\mathcal{I}(e))$. Then we find the heaviest path from r to the genes in M using a simple shortest-path algorithm for weighted graphs. If the score of the complete model is improved, this step will be taken.

This principle is applied in all search steps: The algorithm searches iteratively for the best structure of the module network and the best assignment of genes to modules, while each local step is done both in the module network and in the interaction map. Thus, in *both* dimensions a new hypothesis is inferred from the observations.

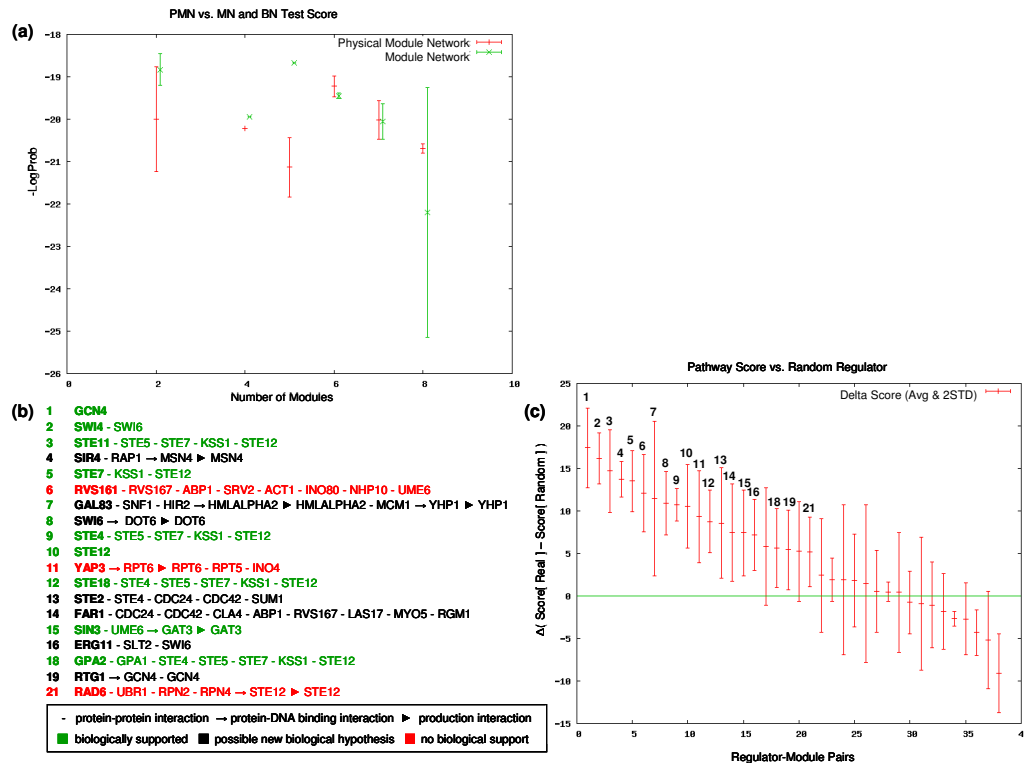


Figure 2: (a) Log Likelihood of the learned networks, for a changing number of modules. Average and 2 standard deviation were estimated from 5-fold training. (b) 19 statistically significant regulation pathways that were reconstructed between regulators and their true target modules. (c) Distribution of differences between the score of the reconstructed pathway and the scores of random pathways. The random pathways were reconstructed with the same module, but with randomly chosen regulators. The statistically significant pathways are numbered.

4 Results and Discussion

Our objectives now are to compare the performance of a PMN with a regular Module Network, and to validate the learned model statistically and biologically.

To evaluate the models, we compare their ability to generalize to unseen data. Both models were trained on a set of gene expression samples, measured from 454 genes in *S. cerevisiae* (Gasch et al. [2000]), where the physical interaction observations were collected from different biological and computational sources (Harbison et al. [2004], Siddharthan et al. [2005], von Mering et al. [2005], Reguly et al. [2006]). The scores of the learned networks were measured with a separate sample test set for a changing number of modules. The results show that the generalization abilities of the models are similar (Figure 2 (a)), and there is no single model that clearly outperforms the other. This could result from the discretization of the expression data, which was performed due to efficiency considerations, but may alter the biological signal. Further examination with continuous data may distinguish PMNs from Module Networks.

Even though the PMN did not generalize better than the Module Networks, it still provides new biological information in the form of regulation pathways. Thus, we next examined whether the reconstructed pathways are biologically and statistically meaningful. To test the model with a sound set of regulators and their true biological target genes, we used a set of expression profiles measured in a series of knock-outs in *S. cerevisiae* (Hughes et al. [2000]). Each knocked-out gene was defined as a regulator, and the down-regulated genes in the corresponding experiment were the module genes. For each such regulator-module pair, the regulation pathway was reconstructed, and its score was compared to a score distribution, obtained with the same module but randomly chosen regulators (Figure 2 (c)). Among the 45 pathways, 19 were statistically significant, where 9 are biologically meaningful, 3 has no biological support, and 7 present a potentially new hypothesis that needs further investigation (Figure 2 (b)). These encouraging results demonstrate that the model can predict the mechanisms by which a gene knock-out results in large scale gene expression changes.

To conclude, the integration of large scale data from different sources into a fully stochastic framework suggests a promising approach to the reconstruction of regulatory mechanisms. Here we presented a model that integrates gene expression and physical interactions data, and this concept can be extended to include other types of molecular data.

References

- N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *J. Comp. Bio.*, 7:601–620, 2000.
- A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. Genomic expression program in the response of yeast cells to environmental changes. *Mol. Bio. Cell*, 11:4241–4257, 2000.
- CT. Harbison, DB. Gordon, TI. Lee, NJ. Rinaldi, KD. Macisaac, TW. Danford, NM. Hannett, JB. Tagne, DB. Reynolds, J. Yoo, EG. Jennings, J. Zeitlinger, DK. Pokholok, M. Kellis, PA. Rolfe, KT. Takusagawa, ES. Lander, DK. Gifford, E. Fraenkel, and RA. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104, 2004.
- A. Hartemink, D. Gifford, T. S. Jaakkola, and R. Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In *Pac. Symp. Biocomp.* 6, 2001.
- T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepanians, D. D. Shoemaker, D. Gachotte, K. Chakraburty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–26, 2000.
- T. Ideker, O. Ozier, B. Schwikowski, and A.F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18:S233–S240, 2002.
- N. Nariai, Y. Tamada, S. Imoto, and S. Miyano. Estimating gene regulatory networks and protein-protein interactions of saccharomyces cerevisiae from multiple genome-wide data. *Bioinformatics*, 21; Suppl 2:ii206–ii212, 2005.
- T. Reguly, A. Breitkreutz, L. Boucher, B.J. Breitkreutz, G.C. Hon, C.L. Myers, A. Parsons, H. Friesen, R. Oughtred, A. Tong, C. Stark, Y. Ho, D. Botstein, B. Andrews, C. Boone, O.G. Troyanskaya, T. Ideker, K. Dolinski, N.N. Batada, and M. Tyers. Comprehensive curation and analysis of global interaction networks in saccharomyces cerevisiae. *J Biol.*, 2006.
- E. Segal, Y. Barahs, I. Simon, N. Friedman, and D. Koller. From promoter sequence to expression: a probabilistic framework. In *RECOMB '02*. 2002.
- E. Segal, D. Peer, A. Regev, D. Koller, and N. Friedman. Learning module networks. *JMLR*, 6:557–88, 2005.
- E. Segal, M. Shapira, A. Regev, D. Peer, D. Botstein, D. Koller, and N. Friedman. Modules networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166–76, 2003a.
- E. Segal, H. Wang, and D. Koller. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19: 264–272, 2003b.
- R. Siddharthan, ED. Siggia, and E. van Nimwegen. Phylogibbs: A gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol*, 1, 2005.
- C. von Mering, LJ. Jensen, B. Snel, SD. Hooper, M. Krupp, M. Foglierini, N. Jouffre, MA. Huynen, and P. Bork. String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*, pages D433–7, 2005.
- C.H. Yeang, T. Ideker, and T. Jaakkola. physical network model. *JCB*, 11:243–62, 2004.