INFERRING GENETIC REGULATORY INTERACTIONS FROM TIME-COLLAPSED BOOLEAN SUMMARY VARIABLES

Timo Erkkilä[1,2]*, Vesteinn Thorsson*[2]*, Harri Lähdesmäki*[1,3]*, and Ilya Shmulevich*[2,1]

[1]Department of Signal Processing, Tampere University of Technology, Finland
[2]Institute for Systems Biology, Seattle, WA, USA
[3]Department of Information and Computer Science, Helsinki University of Technology, Finland
{timo.p.erkkila,harri.lahdesmaki}@tut.fi, {vthorsson,ishmulevich}@systemsbiology.org

ABSTRACT

Inference of causal relationships from gene expression data sets in order to identify possible modes of gene regulation is arguably one of the key problems in computational systems biology. However, not much attention has been paid to how the ever-increasing amount of data for inference will be handled in a feasible manner. We tackle this data expansion problem by proposing a probabilistic Boolean logic model for steady-state and time-course gene expression data that accounts for large data sets. Furthermore, ranking of genes with a novel measure of importance is proposed.

1. INTRODUCTION

With the growing availability of large-scale and high-throughput experimental assays, methods for identifying relationships between genes are becoming increasingly important. However, researchers are frequently faced with a situation in which the assays can differ greatly, for instance, in frequency of time sampling and the experimental platforms used. To be able to make use of all available data in such situations, there is a need for methods that characterize the observations in a robust manner, independent of aforementioned differences in experimental design, and draw interesting relationships. In addition, such methods should cope with large numbers of genes and experiments.

Here we present a Probabilistic Boolean Network (PBN) [1] inference framework in which relationships among genes of interest are inferred from multiple gene expression experiments consisting either of steady-state or time-course measurements, or both. For summarizing the strengths of interactions between genes of interest over multiple network models, we use the so-called influence of a gene on another gene [2]. Furthermore, we propose a quantity called *importance* with which ranking of genes in a given influence network can be performed. Genes serving as strong influence hubs (multiple incoming and outgoing influence edges) in the network will receive large importance, and vice versa.

We utilize our framework with a publicly available data set that consists of time-course gene expression profiles of bone marrow-derived macrophages (BMDMs), stimulated by a variety of pathogen-associated molecular patterns (PAMPs). We focused our analysis on cytokines and transcription factors, functional groups that play important roles in immune responses: cytokines are communication molecules used by immune cells and transcription factors are key players in the regulatory networks governing the immune response. After data pre-processing, our interest focused on 170 genes, falling into these two functional categories, that were transcriptionally activated by PAMPs, followed by further analysis such as network visualization and gene set enrichment analysis (GSEA) [3] with the computed influences; importance is utilized for providing a ranked gene list for GSEA.

2. METHODS

2.1. Constructing the PBN

At this point we assume that Boolean summary variables for N genes under E experiments have been extracted and stored in $X \in \{0,1\}^{N\times E}$; an example derivation of Boolean summary variables for the immunity data will be given in Sec. 3. With X we construct a set of functions, $\mathcal{F}_n$, for gene (node) $n \in \{1,2,\ldots,N\}$, where to each function $f_{n,i} \in \mathcal{F}_n$ a selection probability $s_{n,i} \in \mathcal{S}_n$ is associated. Our inference algorithm uses the best-fit criterion [4], whereas for having model complexity embedded into the framework, the minimum description length (MDL) [5] principle will be used to control model complexity.

Upon splitting the function space into three sub-categories according to the number of significant inputs, $m \in \{0,1,2\}$, we can then search for the best-fit function $f_{n,i}^{(m)}$ for each gene n and input combination i with the PBN toolbox `http://personal.systemsbiology.net/ilya/PBN/PBN.htm` for Matlab (MathWorks, Natick, MA) $f_{n,i}^{(m)} = \mathrm{BF}\left(\mathbf{y}_n, X_n, \mathcal{C}_n(i), \mathcal{A}_m\right)$ for the input data X_n, output data $\mathbf{y}_n$, input pair $\mathcal{C}_n(i)$ (out of $I = \binom{N-1}{2}$ combinations), and Boolean functions $\mathcal{A}_m$ with m significant inputs. The estimated function is associated with the best-fit error, $\epsilon(f_{n,i}^{(m)})$, which equals to the number of misclassifications and which we use for MDL estimation. For each best-fit function, MDL becomes computed with [5]

$$\mathrm{MDL}\left(f_{n,i}^{(m)}\right) = 2^m + \log_2(E)\epsilon(f_{n,i}^{(m)}) + 2, \quad (1)$$

a mapping that balances with model complexity (first summand) and prediction error (second summand). For each triplet $\{f_{n,i}^{(0)}, f_{n,i}^{(1)}, f_{n,i}^{(2)}\}$, we select that one into $\mathcal{F}_n$ that has the smallest MDL:

$$f_{n,i} = \arg \min_{f_{n,i}^{(m)}:m\in\{0,1,2\}} \left\{\mathrm{MDL}\left(f_{n,i}^{(m)}\right)\right\}. \quad (2)$$

Each function $f_{n,i} \in \mathcal{F}_n$ is now associated with an MDL, $\mathrm{MDL}\left(f_{n,i}\right)$, which gives the minimum number of bits to represent the data [5]. Thus, an increase of MDL by one means that *at least* two times more information is required for coding. Furthermore, existing theoretical work links minimum code-lengths, i.e., MDLs, to probability measures by the mapping [6]

$$\mathbb{P}\left[\mathrm{MDL}\left(f_{n,i}\right)\right] \propto 2^{-\mathrm{MDL}(f_{n,i})}, \quad (3)$$

which basically states the same observation analogously with probabilities: an increase of one bit in function coding decreases the probability of that code-length by a factor of two. We exploit that result by imposing $s_{n,i} = \mathbb{P}\left[\mathrm{MDL}\left(f_{n,i}\right)\right]$, i.e., the selection probabilities directly follow the distribution given by Eq. (3).

2.2. Computing influence

The function sets $\mathcal{F}_1,\ldots,\mathcal{F}_N$ and selection probabilities $\mathcal{S}_1,\ldots,\mathcal{S}_N$ are a representation of the dependencies across genes in the PBN. To extract salient network features shared by the multiple possible networks, we computed a pairwise measure called influence [2], which quantifies the ability of one gene to affect a change in the level of a gene it regulates directly. As a result we have a $G \in [0,1]^{N\times N}$ square matrix whose elements are the influences ($g_{i,j}$ is the influence from gene i to gene j).

2.3. Gene importance

To identify which genes are essential in the network, i.e., those having multiple incoming and outgoing, strong influences across genes, a summary statistic called *importance* is proposed. With such a measure it is possible to computationally identify strong members within cliques as well as cliques themselves, and rank genes based on the measure of importance; more important genes will be ranked higher, and this ranked list could be used for, say, GSEA [3]. Importance of gene n, Ψ_n, takes into account the aforementioned features in the following, additive fashion: $\Psi_n = \alpha_n + \beta_n + \gamma_n + \delta_n$; α_n is the standardized sum of incoming influences; β_n is the standardized sum of outgoing influences; γ_n is the standardized maximum incoming influence; δ_n is the standardized maximum outgoing influence.

3. RESULTS

3.1. Data pre-processing

The time-course gene expression profiles of BMDMs were obtained from `http://www.`

systemsimmunology.org, and were then normalized and grouped into biological replicates to obtain median values for each distinct biological conditions. Arrays were organized into a total of $E = 49$ time-courses of PAMP stimulation, each for a distinct cell type (BMDMs), strain and PAMP stimulus. Values from PAMP stimulated cells were compared with un-stimulated values to yield ratios of expression induction. Data were discretized such that each time-course was assigned the value 1 if there was a time point with ratio greater than 3 (induction only, not repression) and intensity exceeding 300, and was otherwise assigned the value 0. Genes were restricted to those with Gene Ontology (http://www.geneontology.org) molecular function assignments of "Cytokine Activity" (GO:0005125), and molecular function "Transcription Factory Activity" (GO:0003700), as obtained from the Mouse Genome Informatics http://www.informatics.jax.org. A final filter was applied to discard genes or conditions in which no induction was observed, yielding the final binary matrix $X \in \{0,1\}^{N \times E}$ with $N = 170$ genes and $E = 49$ time-courses.

3.2. Analysis

The analysis for the 170 genes was completed in less than an hour. As a preliminary result, we observed that the frequency with which we observe an inferred Boolean function with m active inputs decreases as m increases (top-left histogram in Fig. 1), and not many of the 170^2 gene-gene interactions were associated with high influences (bottom-left histogram in Fig. 1), supporting the assumption that regulatory connectivity among genes is relatively sparse. Histograms of gene-specific, average incoming and outgoing influences (top-right histogram in Fig. 1) suggest that genes tend to be equally influenced (peak around 0.06), whereas outgoing influences tend to concentrate on values less than 0.06 (not many genes are influencers). Histograms of gene importances (bottom-right histogram in Fig. 1) show that both transcription factors and cytokines are equally "important" in the network.

3.2.1. GSEA

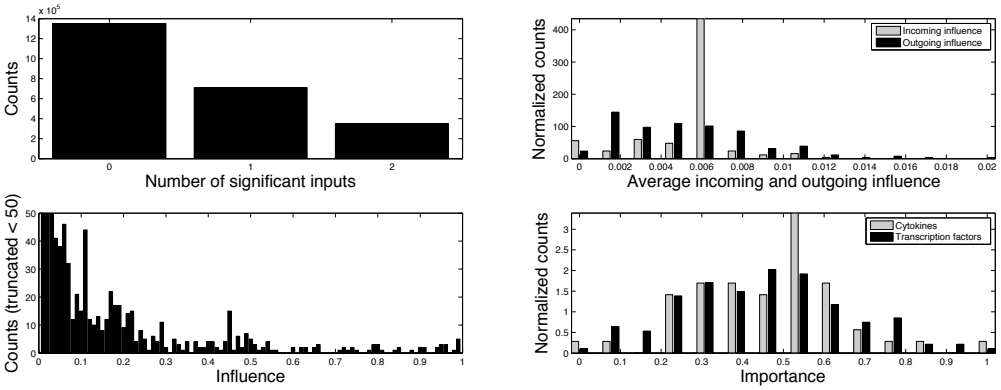
In addition to more or less qualitative analysis with gene influences and importances, we searched for enriched KEGG pathways and GO annotations with GSEA. Genes were first ranked based on their importance, after which enrichment was calculated with the existing web-application, "GeneTrail" (http://genetrail.bioinf.uni-sb.de/). The most highly enriched pathway was toll-like receptor (TLR) signaling pathway ($p = 0.0184847$), being the key link in the activation of immune response to pathogens – here PAMPs. Furthermore, multiple GO categories related to both T-cell activation and positive regulation of transcription were observed ($p < 0.05$).

3.3. Visualization

Visualizing the inferred influence network is helpful for exploring individual hypotheses as well as identifying which possible network structures, e.g., cliques, should be targeted for further computational exploration. Considering only influence edges exceeding an arbitrary threshold 0.2 (for reducing connectivity), we created a Cytoscape [7] visualization, in which influence and importance control edge width and node size, respectively (graph downloadable and viewable at http://www.cs.tut.fi/~erkkila2/ under "Research" section).

4. CONCLUSION

With the proposed PBN framework, analysis of large gene-sets in association with large number of experimental conditions is made possible in a feasible computation time. In order to account for multiple possible network structures, we used the notion of influence for summarizing pairwise interactions across genes from the inferred PBN, and further used the influences for graph visualization purposes. An algorithmic way of ranking genes in a given network structure was proposed, using a ranking statistic termed *importance*. As a proof-of-concept, our method identified relevant GO categories and KEGG pathways significantly enriched.



(a) (b)

Figure 1. Statistics from the inference with PBNs. **Top-left**: Histogram of active inputs in the PBN. **Bottom-left**: histogram of influences; the plot was truncated due to large frequency of small influences). **Top-right**: histogram of average incoming and outgoing influences. **Bottom-right**: histogram of gene importance.

5. ACKNOWLEDGMENTS

TE and HL was supported by the Academy of Finland (applications 121830, Postdoctoral Researcher's Project 2008-2010, application 213462, Finnish Programme for Centres of Excellence in Research 2006 − 2011, and application 134290) and Finnish Graduate School in Computational Sciences (FICS), VT and IS was supported in part by the Contract "Systems Approach to Immunity and Inflammation" (HHSN272200700038C) from the National Institute of Allergy and Infectious Diseases.

6. REFERENCES

[1] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, "Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks.," *Bioinformatics*, vol. 18, no. 2, pp. 261–274, Feb 2002.

[2] I. Shmulevich, E. R. Dougherty, and W. Zhang, "From boolean to probabilistic boolean networks as models of genetic regulatory networks," *In Proceedings of the IEEE*, vol. 90, 2002.

[3] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.," *Proc Natl Acad Sci U S A*, vol. 102, no. 43, pp. 15545–15550, Oct 2005.

[4] H. Lähdesmäki, I. Shmulevich, and O. Yli-Harja, "On learning gene regulatory networks under the boolean network model," *Machine Learning*, vol. 52, pp. 147–167, 2003.

[5] I. Tabus and J. Astola, "On the use of mdl principle in gene expression prediction," *EURASIP Journal on Applied Signal Processing*, vol. 2001, pp. 297–303, 2001.

[6] J. Rissanen, *Information and Complexity in Statistical Modeling*, Springer-Verlag New York, LLC, 2007.

[7] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks.," *Genome Res*, vol. 13, no. 11, pp. 2498–2504, Nov 2003.