Computing Minimal Boolean Models of Gene Regulatory Networks

Guy Karlebach,[1*] ,Peter N Robinson[1,2]
[1]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA,
[2]Institute for Systems Genomics, University of Connecticut, Farmington, CT, USA
[*]To whom correspondence should be addressed; E-mail: guy.karlebach@jax.org.

***Abstract*—Objective: Models of Gene Regulatory Networks (GRNs) capture the dynamics of the regulatory processes that occur within the cell as a means to understand the variability observed in gene expression between different conditions. Possibly the simplest mathematical construct used for modeling is the Boolean network, which dictates a set of logical rules for transition between states described as Boolean vectors. Due to the complexity of gene regulation and the limitations of experimental technologies, in most cases knowledge about regulatory interactions and Boolean states is partial. In addition, the logical rules themselves are not known a-priori. Our goal in this work is to present a methodology for inferring this information from the data, and to provide a measure for comparing network states under different biological conditions. Methods: We present a novel methodology for integrating experimental data and performing a search for the optimal consistent structure via optimization of a linear objective function under a set of linear constraints. We also present a statistical approach for testing the similarity of network states under different conditions. Results: Our methodology finds the optimal model using an experimental gene expression dataset from human CD4 T-cells and shows that network states are different between healthy controls and rheumatoid arthritis patients. Conclusion: The problem can be solved optimally using real-world data. Properties of the inferred network show the importance of a general approach. Significance: Our methodology will enable researchers to obtain a better understanding of the function of gene regulatory networks and their biological role.**

I. Introduction

Maintenance of cellular functions requires the orchestration of many interleaving processes over time and space. A Gene Regulatory Network (GRN) is a set of genes such that the present state of their expression trajectory can be predicted from past states via the regulatory relationships between the genes. Due to their simple building blocks GRNs have been used to model the regulation of processe as different as cell differentiation, circadian clocks and diauxic shift [1], [2], [3]. Consequently, many methods for reconstructing GRNs from experimental data at varying levels of details have been proposed [4], [5], [6]. The simplest formulation, the Boolean network, describes gene expression states as Boolean values and changes in those levels as Boolean functions [7]. While the simplicity of this model imposes a certain level of abstraction it also makes it applicable to a broader range of datasets and simplifies its analysis. Interestingly, despite the relative simplicity of Boolean networks, fitting a Boolean network to a gene expression dataset given a set of regulatory relationships is an NP-Hard problem [8]. In practice the exact regulatory relationships are not known, which can result in redundant regulators after fitting. A possible remedy to this problem is impose non-redundant logic as a requirement in the solution of the fitting problem [4], [9]. However, this approach contradicts the widely accepted principle that a simpler model that provides the same degree of fit to the data is preferable to a more complex one [10], [11], [12]. In this paper we present a novel algorithm for fitting a Boolean network to a gene expression dataset that addresses the problem of redundant regulators. In addition, we provide a method to compute the statistical significance of difference between network states under different conditions. We apply our methodology to a gene expression dataset from human CD4 T-cells that was obtained from rheumatoid arthritis patients and healthy controls .

II. Methods

A. Optimal Fit of Boolean Networks

A gene expression dataset consists of a N × M matrix where N corresponds to the number of genes whose expression level was measured and M corresponds to the number of experiments. In a typical dataset N $\gg$ M. The expression values in the matrix can be discrete or continuous, depending on the experimental technology that was used for generating the data, but higher values always represent a higher expression level. In order to map these values to Boolean values one needs to label each observation as belonging to a state of low or high expression. Since the proposed methodology is independent of the choice of a mapping, in the rest of this section we will assume that the mapping has already been applied to the data.

A trajectory of a Boolean network is a sequence of states such that each state except for the first state in the sequence is derived from the previous state using a set of Boolean functions. Each Boolean function determines the value of exactly one gene, and its inputs are the states of any number of genes in the network. Usually, it is assumed that the number of inputs is small compared to the total number of genes. The regulatory relationships of a Boolean network can be illustrated as edges from inputs to outputs in a directed graph, called a regulation graph. A gene that has an outgoing edge to another gene is referred to as a regulator, and a gene with an incoming edge as a target (a gene can be both a regulator and a target). A steady state is a state that repeats itself in a trajectory indefinitely unless perturbed by external signals, i.e. signals that are not part of the network. In a typical gene expression dataset the experiments correspond to a single time point, and therefore the network is assumed to be in a steady state in each experiment. For simplicity of description we assume in the rest of this section that the network is in steady state, however the algorithm presented here is applicable to any type of data.

Discrepancies between a dataset and a network model occur when the Boolean values in an experimental dataset do not agree with any network trajectory due to experimental noise. This presents a difficulty if the network model is not known a-priori, since enumerating all possible networks is infeasible. Formally, let $C_{g_i,j}$ denote the Boolean value of gene g_i in experiment j, and let e_{g_1,g_2} denote a directed edge between genes g_1 and g_2. We say that the data contains a discrepancy if for some gene g and two experiments i_1 and i_2, $C_{g_j,i_1} = C_{g_j,i_2}$ for all genes g_j such that an edge $e_{g_j,g}$ exists, but $C_{g,i_1} \neq C_{g,i_2}$. It follows from the network's determinism that at least one of the experiments i_1 or i_2 does not agree with any network trajectory.

Assuming that $P \neq NP$, there does not exist a polynomial time algorithm for resolving all the discrepancies with the minimal number of changes. Therefore, either a heuristic that finds a local optimum or an algorithm that may not terminate in a reasonable amount of time must be used instead. Another difficulty is that a strict subset of the regulation graph may provide a solution with the same number of changes to the expression dataset, and so the structure of the network itself needs to be considered in the search for the optimal solution. This brings another level of complexity to the already difficult problem.

B. An Algorithm for the Optimal Minimal Network

The in-degree of nodes in the regulation graph is usually assumed to be small compared to the number of genes or the number of experiments. If we assume that it is a constant value in terms of computational complexity, we can define a set of constraints on the values that have to be changed in order to remove all discrepancies from the data. Let C_{ij} denote the Boolean input value of gene i at experiment j, and let $B_{g_i,j}$ equal 1 if the value of gene g_i in experiment j was flipped in the solution, and otherwise 0. Then for every experiment j and for every gene g_{k+1} with regulators $g_1, g_2, ..., g_k$ and for every Boolean vector $(w_1, w_2, ..., w_k)$, $w_j \in \{0,1\}$, if the output of the Boolean function that determines the value of g_{k+1}, $I(w_1, w_2, ..., w_k)$, is 1, the following constraint must hold:

$$\begin{gathered}\sum_{r=1}^{k}(C_{rj}\cdot(w_r+(1-2\cdot w_r)\cdot B_{g_r,j}) \qquad (1)\\ +(1-C_{rj})\cdot((1-w_r)+(2\cdot w_r-1)\cdot B_{g_r,j}))\\ +C_{k+1}\cdot B_{g_i,j}+(1-C_{k+1})\cdot(1-B_{g_i,j})\\ <(2-I(w_1,w_2,...,w_k))\cdot(k+1)\end{gathered}$$

This constraint means that if the output variable $I(w_1, w_2, ..., w_k)$ was set to 1, whenever the inputs $w_1, w_2, ..., w_k$ appear in the solution the output (the value of g_{k+1}) must be 1. Similarly, if $I(w_1, w_2, ..., w_k)$ is set to 0 the following constraint must hold:

$$\begin{gathered}\sum_{r=1}^{k}(C_{rj}\cdot(w_r+(1-2\cdot w_r)\cdot B_{g_r,j}) \qquad (2)\\ +(1-C_{rj})\cdot((1-w_r)+(2\cdot w_r-1)\cdot B_{g_r,j}))\\ +C_{k+1}\cdot(1-B_{g_i,j})+(1-C_{k+1})\cdot B_{g_i,j}\\ <(I(w_1,w_2,...,w_k)+1)\cdot(k+1)\end{gathered}$$

By requiring that under these constraints the following sum is minimized:

$$\sum\nolimits_{j\in 1,..,M}^{i\in 1,..,N} B_{ij}$$

we can use a branch and bound algorithm for 0/1 integer programming to find a solution that fits the data with a minimal number of changes and construct a new dataset with values $D_{ij} = ((C_{ij} + B_{ij}) \mod 2), i \in 1..N, j \in 1..M$. However, this formulation still ignores the possible existence of multiple optimal solutions that correspond to different network structures, for if after resolution of the discrepancies only a strict subset of inputs uniquely determines a function's output then the edges in the regulation graph that correspond to the rest of the inputs can be removed. In order to choose the solution that results in the smallest network, for every gene g_i and each one of its targets g_j, we create another Boolean variable R_{ij}, that is added to the right hand side of additional constraints created for g_j. These constraints are similar to (1) and (2) but with all subsets of regulators removed and with their R_{ij} variables in the r.h.s. If the R_{ij} are equal to 1 then given a solution that satisfies the the full constraints, the new constraints will be satisfied as well. If the new constraints can be satisfied without setting R_{ij} to 1, then the edge from g_i to g_j in the regulation graph is redundant. Therefore, all the variables R_{ij} have weight $\frac{1}{|E|+1}$ in the objective function, where $|E|$ is the number of edges in the regulation graph. The number of variables in the resulting 0/1 integer linear programming is $M \cdot N + \sum_{i=1}^{N} 2^{indegree(g_i)} + |E|$. Once the corrected values D_{ij} are obtained from the optimal solution, we remove redundant regulators to obtain the inferred network structure.

Given the optimal network and its inferred states, we now wish to compute the probability of seeing the observed difference between two groups of states, referred to as cases and controls, by chance. In order to do that, we compute the difference between the expected between-group state distance and the expected within-group state distance, where the two groups are cases and controls. A distance between two binary states is simply the sum of non-identical entries.

$$\frac{\sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M}\sum_{k=1}^{N}|D_{ki}-D_{kj}|}{M\cdot(M-1)}- \quad (3)$$

$$\frac{\sum_{i\in group1}\sum_{j\in group1,j\neq i}\sum_{k=1}^{N}|D_{ki}-D_{kj}|}{|group1|\cdot|group1-1|\cdot 2}-$$

$$\frac{\sum_{i\in group2}\sum_{j\in group2,j\neq i}\sum_{k=1}^{N}|D_{ki}-D_{kj}|}{|group2|\cdot|group2-1|\cdot 2}$$

In order to generate a value from the null distribution we sample a number of random states equal to the number of states in the data and apply the network logic to find steady states. We then compute the same statistic (3) by randomly assigning the patient and healthy labels from the data to these states.

III. Results

A. A Gene Regulatory Network for T-cell Migration

In order to test our algorithm we obtained gene expression values from the dataset of Ye et al. [13] and constructed a regulation graph using transcription factor-targets pairs from the ORegAnno database [14]. Ye et al. [13] measured gene expression in CD4+ T cells from 13 rheumatoid arthritis patients and 9 healthy controls. After RMA normalization, we retrieved genes corresponding to the Gene Ontology term ”leukocyte tethering or rolling” (GO:0050901) and their regulators from ORegAnno [14]. For each gene, expression values greater or equal to the mean expression level were mapped to a binary 1, and expression values lower than the mean to a binary 0. A set of transcription factors such that the conditional entropy of their target is at most 0.2 was selected for each gene using forward selection. This corresponds to a fraction of 0.05 target values that are not predicted by the regulators. The resulting network contained 57 genes, from which 29 are transcription factors and 28 are targets. In order to fit the minimal network to the gene expression data, we used the Gurobi solver [15] with parameter GomoryPasses=0. The optimal solution flips 54 Boolean values, which corresponds to a noise level of approximately 4.3%.

B. Analysis of the Minimal Network's Structure and Logic

Figure 1 displays a multidimensional scaling of the samples using the binary distance between them after network fitting. The figure shows that the patient and healthy samples are linearly separable and that there is more diversity among patient samples than among healthy controls. After fitting, 9 of the 28 targets were shown to have had a redundant regulator that was removed in the minimal network. Figure 2 shows the number of regulators of each gene before and after the fitting. We conclude that fitting improves our ability to assign regulators to targets.

Since the enzymes encoded by the genes FUT4, FUT7 and FUT9 play a role in the production of the Lewis antigen whose presence on leukocytes has been linked to immune disorders, we examined the corresponding regulation functions in order to determine which type of intervention would be needed to down-regulate the expression of the target genes. The minimal interventions that down-regulate the transcription of FUT9 are either up-regulating CTCF, down-regulating FOXA1 or down-regulating TP53, depending on the patient. Down-regulating of FUT4 expression requires down-regulation of either ESR1 or

EGR1, depending on the patient. In some of the patients, down-regulating the transcription of FUT7 requires modifying the expression or activity of at least two transcription factors, and therefore it is the least favorable target with respect to perturbation size.

In order to characterize the complexity of the regulatory logic we examined for all genes the correlation between the number of expressed regulators and the expression of the target. The mean Kendall's tau was 0.03, indicating that the output of regulation functions is in general not associated with the number of expressed regulators. Furthermore, the correlation was not associated with the number of regulators.

C. Comparison of Patient and Healthy Network States

The noise level of different genes in the model is of interest since assumptions about this parameter are often made in models of gene expression. Therefore, we examined for each gene the number of input values of each type (Boolean 0 and 1) that were flipped in the optimal fit. The distribution of noise across the genes was not symmetric and was not associated with an expressed state (Boolean 1) or non-expressed state(Boolean 0). This suggests that modeling assumptions that assign an equal level of noise for all genes may lead to wrong conclusions. In order to test the hypothesis that the group of patient network states and the group of healthy network states come from the same distribution, we calculated the difference between the expected between-group state distance and the expected within-group distance, and generated the null distribution as described methods section. This procedure does not make assumptions about the distribution of measurement noise. The p-value obtained by calculating the statistic for 100 permutations of the data was less than 0.01. The null distribution and the value of the statistic calculated for the data are displayed in Figure 3. We conclude that patient and network states are likely generated from different distributions.

IV. Conclusion

We propose a new algorithm for fitting a Boolean network model to gene expression data that improves on previous approaches by minimizing the size of the network that fits the data optimally. Using a database of curated transcription factor-target pairs and steady-state data from rheumatoid arthritis patients and healthy controls, we found the minimal network structure that fits the data optimally. Inspection of the structural properties of the inferred network show that fitting prunes redundant regulators, which stresses the importance of sparsity constraints in the search algorithm. Furthermore, the inferred logic was diverse and showed that constraints on the logic functions should be avoided. By sampling random states and applying the inferred network logic we found that patient and healthy network states are likely to arise from different distributions. The regulatory relationships between transcription factors and their targets as given by the model and the expression profiles of the patients enabled the selection of transcription factors that are candidates for therapeutic intervention. Changing the level of these factors in affected tissues, according to the model, will reduce the expression of genes that are essential for leukocyte migration into the area of inflammation.

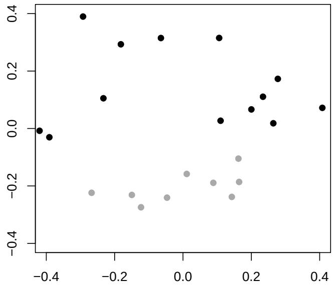

Fig. 1: Multidimensional scaling of fitted samples

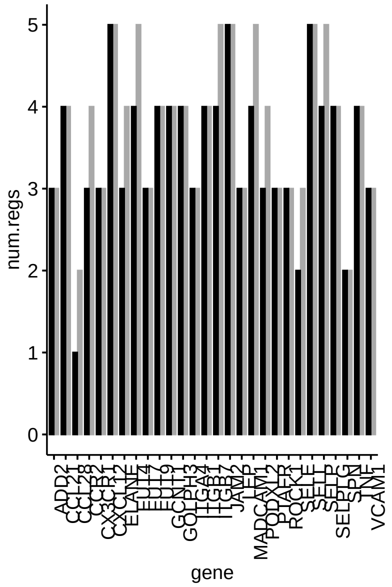

Fig. 2: Number of regulators before and after fitting

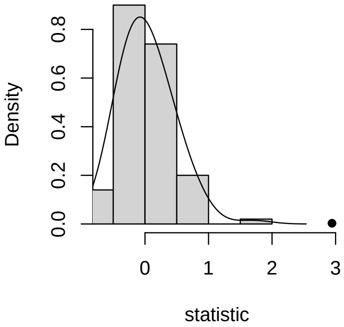

Fig. 3: Patient-healthy state difference statistic and null distribution

References

[1] L. Geistlinger, G. Csaba, S. Dirmeier, R. Kuffner, and R. Zimmer, “A comprehensive gene regulatory network for the diauxic shift in saccharomyces cerevisiae,” *Nucleic Acids Research*, vol. 41, no. 18, pp. 8452–8463, Jul. 2013. [Online]. Available: https://doi.org/10.1093/nar/gkt631

[2] T. Ohara, T. J. Hearn, A. A. Webb, and A. Satake, “Gene regulatory network models in response to sugars in the plant circadian system,” *Journal of Theoretical Biology*, vol. 457, pp. 137–151, Nov. 2018. [Online]. Available: https://doi.org/10.1016/j.jtbi.2018.08.020

[3] S. N. Willis and S. L. Nutt, “New players in the gene regulatory network controlling late b cell differentiation,” *Current Opinion in Immunology*, vol. 58, pp. 68–74, Jun. 2019. [Online]. Available: https://doi.org/10.1016/j.coi.2019.04.007

[4] G. Karlebach and R. Shamir, “Modelling and analysis of gene regulatory networks,” *Nature Reviews Molecular Cell Biology*, vol. 9, no. 10, pp. 770–780, Sep. 2008. [Online]. Available: https://doi.org/10.1038/nrm2503

[5] Y. Shavit, B. Yordanov, S.-J. Dunn, C. M. Wintersteiger, T. Otani, Y. Hamadi, F. J. Livesey, and H. Kugler, “Automated synthesis and analysis of switching gene regulatory networks,” *Biosystems*, vol. 146, pp. 26–34, Aug. 2016. [Online]. Available: https://doi.org/10.1016/j.biosystems.2016.03.012

[6] R. F. Hashimoto, S. Kim, I. Shmulevich, W. Zhang, M. L. Bittner, and E. R. Dougherty, “Growing genetic regulatory networks from seed genes,” *Bioinformatics*, vol. 20, no. 8, pp. 1241–1247, Feb. 2004. [Online]. Available: https://doi.org/10.1093/bioinformatics/bth074

[7] S. Kauffman, “Metabolic stability and epigenesis in randomly constructed genetic nets,” *Journal of Theoretical Biology*, vol. 22, no. 3, pp. 437–467, Mar. 1969. [Online]. Available: https://doi.org/10.1016/0022-5193(69)90015-0

[8] G. Karlebach and R. Shamir, “Constructing logical models of gene regulatory networks by integrating transcription factor dna interactions with expression data: An entropy-based approach,” *Journal of Computational Biology*, vol. 19, no. 1, pp. 30–41, Jan. 2012. [Online]. Available: https://doi.org/10.1089/cmb.2011.0100

[9] R. Sharan and R. M. Karp, "Reconstructing boolean models of signaling," *Journal of Computational Biology*, vol. 20, no. 3, pp. 249–257, Mar. 2013. [Online]. Available: https://doi.org/10.1089/cmb.2012.0241

[10] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, Mar. 1978. [Online]. Available: https://doi.org/10.1214/aos/1176344136

[11] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Springer Series in Statistics*. Springer New York, 1998, pp. 199–213. [Online]. Available: https://doi.org/10.1007/978-1-4612-1694-015

[12] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *The Annals of Statistics*, vol. 11, no. 2, Jun. 1983. [Online]. Available: https://doi.org/10.1214/aos/1176346150

[13] H. Ye, J. Zhang, J. Wang, Y. Gao, Y. Du, C. Li, M. Deng, J. Guo, and Z. Li, "Cd4 t-cell transcriptome analysis reveals aberrant regulation of STAT3 and wnt signaling pathways in rheumatoid arthritis: evidence from a case control study," *Arthritis Research and Therapy*, vol. 17, no. 1, Mar. 2015. [Online]. Available: https://doi.org/10.1186/s13075-015-0590-9

[14] S. B. Montgomery, O. L. Griffith, M. C. Sleumer, C. M. Bergman, M. Bilenky, E. D. Pleasance, Y. Prychyna, X. Zhang, and S. J. M. Jones, "Oreganno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation," *Bioinformatics*, vol. 22, no. 5, pp. 637–640, Jan. 2006. [Online]. Available: https://doi.org/10.1093/bioinformatics/btk027

[15] L. Gurobi Optimization, "Gurobi optimizer reference manual," 2021. [Online]. Available: http://www.gurobi.com

Status

Submitted to IEEE Transactions on Biomedical Engineering May 11th 2021
Returned without review, reason subject is too specialized May 30th 2021
Submitted to next journal May 30th 2021