EXTRACTION OF TRANSCRIPTION FACTOR NETWORKS VIA GLOBALLY OPTIMAL BICLUSTERING

E. Yang[1], P.T. Foteinou[1], K.R. King[2], M.L. Yarmush[1,2] and I.P Androulakis*[1]
[1]Rutgers University, Department of Biomedical Engineering, Piscataway, NJ 08853
[2] Center for Engineering in Medicine/Surgical Services, Massachusetts General
Hospital, Harvard Medical School, Boston, MA 02114, USA

Abstract

We present an optimization-based framework for identifying and quantifying interactions transcription factor networks. We explore the availability of high-temporal resolution expression data using the Living Cell Array and we formulate to critical problems. First, we demonstrate how to rigorously obtain bi-clusters of transcription factor and condition through a novel MILP formulation and subsequently demonstrate how such networks can be quantify using appropriate deconvolution schemes based on linear programming.

Keywords

Bi-clustering, Transcription Network Reconstruction

Introduction

The Living Cell Array is a micro-fluidics device which utilizes cells transfected with artificially constructed reporter plasmids(King, Wang et al. 2007). These reporter plasmids consist of a minimal promoter and 4 repeats of a transcription factor's consensus sequence as identified via the TRANSFAC database(Matys, Fricke et al. 2003), and an unstable GFP. Therefore, the activation of a given transcription factor is correlated with the fluorescence of the given cell. In this experimental context, the activation of a given transcription factor is performed by utilizing a soluble factor which is known to activate that transcription factor. An example of this would be the use of TNF-α for the activation of NFkB.

One of the questions which we seek to answer is whether, such data is sufficient for the purposes of identifying how the activation of one transcription factor can affect the activity of another. Our starting point is the hypothesis that transcription factors which show similar activity under multiple stimulatory conditions will be linked, and if these links can be fully identified, then it should be possible to construct a network which describes how all of the measured transcription factors interact.

The primary computational problem which must be solved is the identification of a subset of conditions in which a group of transcription factors show similar activation kinetics. This is because most genes are activated by multiple transcription factors, and therefore co-expressed genes may not be co-expressed under all conditions. Finding this set of conditions and genes/transcription factors is the bi-clustering problem(Cheng and Church 2000).

Bi-clustering has been previously identified as being NP-Hard(Zhang 2002), and is normally solved via heuristics. However, the use of heuristic algorithms make it difficult to determine if the largest given bi-cluster has been found. Furthermore, the use of heuristic approaches makes it difficult to add in additional constraints to allow the algorithm to find arbitrarily overlapping bi-clusters(Cano, Adarve et al. 2007). The second issue is more serious in the reconstruction of transcriptional

* Corresponding Author

networks because without overlapping bi-clusters, the data then is broken down into independent sub-networks which runs counter to the hypothesis that biological networks are highly connected(Freeman, Goldovsky et al. 2007).

While there exist various algorithms that can find overlapping bi-clusters(Liu and Wang 2007), they normally require some *a priori* knowledge. By utilizing a math programming formulation, it is possible to obtain bi-clusters which are guaranteed to be optimal as well as arbitrarily overlapping bi-clusters.

Methods

Data

The data obtained from the LCA consists of N transcription factors, M conditions, and t time points **Figure 1**. Therefore, for each combination of transcription factors and activation via soluble factors, a time series is reported rather than a single value. To make use of a bi-clustering formulation, this 3 dimensional data was reduced into a two dimensional data via k-means clusters. For bi-clustering one is interested in whether transcription factors are co-expressed under a given condition. Therefore, the k-means clustering was performed on all of the transcription factors for a given condition, such that if the two transcription factors are co-expressed under a given condition, they are assigned to the same cluster thereby giving them the same cluster index.

The data itself consists of the transcription factors and stimulatory factors given in **Table 1.** The transcription factor HSE is not shown despite being used in the experiment because it did not have a specific activator in the experiment. Additionally, there were stimulatory conditions consisting of Lipopolysaccharide(LPS), Cyts, Cyts+Dex. LPS is an inflammatory endotoxin and for the purposes of this analysis not associated with a specific transcription factor and functions as a general inflammatory signal. Cyts represents a combination of all of the different soluble signals without Dexamethasone, and Cyts+Dex represents stimulation with all of the different factors. These two conditions were excluded because they represent a composite stimulus in which all of the factors could be stimulated at the same time. Additionally, in the data, NT functioned as a negative control whereas D4G functioned as a positive control.

Table 1: Stimulatory Factors and their associated Transcription Factors

Soluble Factors	Transcription Factor
TNF-α	NFkB
Dexamethasone	GRE
IL1	AP1
IL6	STAT3
ISRE	IFN-γ

Bi-Clustering

This bi-clustering formulation assumes that all of the transcription factors under a given condition ought to have the same value or dynamics. This is one of the many ways a bi-cluster can be defined(Madeira and Oliveira 2004).

The identification of a bi-cluster is given in (1).

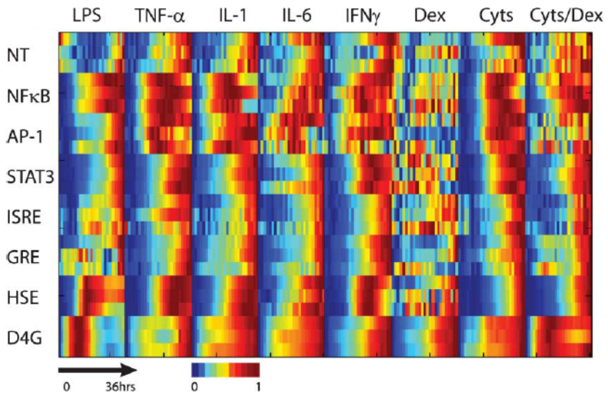

Figure 1: The data obtained form the Living Cell Array . Each column denotes different simulation via a different soluble factor, Each row denotes the response of a given transcription factor(King, Wang et al. 2007).

$$\begin{aligned} \Gamma * [3-(\lambda_i+\lambda_{i'}+\mu_k)] &\geq (\lambda_i+\mu_k)*D(i,k)-(\lambda_{i'}+\mu_k)*D(i,k) \\ \Gamma * [(\lambda_i+\lambda_{i'}+\mu_k)-2] &\leq (\lambda_i+\mu_k)*D(i,k)-(\lambda_{i'}+\mu_k)*D(i,k) \end{aligned} \quad (1)$$

$$i,i' = [1..N]$$
$$k = [1..M]$$

In this formulation, λ represents a binary vector which denotes the assignment of a transcription factor into a bi-cluster, whereas μ represents the assignment of a condition into a bi-cluster. Therefore, if a given transcription factor or condition has been selected, the corresponding binary variable is assigned a one, whereas if it is not selected it is assigned a zero. Γ represents a large number, D the transformed data, with the indices i,i',k denoting the index of transcription factors and conditions.

This formulation attempts to pick up a vector λ and μ such that for all of the selected genes in a given column, the indices returned via the k-means clustering are constant. In addition to these constraints, the objective function is defined as (2) with an additional constraint (3).

$$\max : \sum_{i=1}^{M} \lambda_i \quad (2)$$

$$\sum_{k=1}^{M} \lambda_k = P \quad (3)$$

One of the issues which need to be dealt with each bi-clustering formulation is the definition of the optimum. In the formulation, the goal is to find the largest number of genes which are co-expressed under P different conditions. This formulation is then solved parametrically for all possible values of P [N..1]. This formulation was selected over more commonly used criteria such as maximal area because it is unclear why slightly smaller bi-clusters may be less significant. For example, it is difficult to distinguish between the relative importance of a 10x10 bi-cluster vs. a 8x12 bi-cluster. The important aspect however is that no bi-clustering formulation should find a bi-cluster which is wholly a subset of another previously found bi-cluster.

Constraint (4), is a variation of the Integer-Cuts formulation and allows us to reject solutions that are whole subsets of previously obtained solutions.

$$\begin{aligned} &\sum_{P_{iter}} \mu_k^{citer} - \sum_{Q_{iter}} \mu_k^{citer} < \sum_{k} \mu_k^{citer} \\ &P_{iter} = \{k \mid \mu_k^{iter} = 1\} \\ &Q_{iter} = \{k \mid \mu_k^{iter} = 0\} \end{aligned} \tag{4}$$

The overall bi-clustering formulation is given in (5)

$$\begin{aligned} &\max \sum_{\mathrm{k}} \mu_{\mathrm{k}}{}^{\mathrm{citer}} \\ &\mathrm{s.t} \\ &\sum_{\mathrm{i}} \lambda_{\mathrm{i}}{}^{\mathrm{citer}} = \mathrm{N} \\ &[(\lambda_{\mathrm{i}}{}^{\mathrm{citer}} + \lambda_{\mathrm{i'}}{}^{\mathrm{citer}} + \mu_{\mathrm{k}}{}^{\mathrm{citer}}) - 3] * \Gamma \le (\lambda_{\mathrm{i}}{}^{\mathrm{citer}} + \mu_{\mathrm{k}}{}^{\mathrm{citer}}) * \mathrm{D(i,k)} - (\lambda_{\mathrm{i'}}{}^{\mathrm{citer}} + \mu_{\mathrm{k}}{}^{\mathrm{citer}}) * \mathrm{D(i',k)} \\ &[3 - (\lambda_{\mathrm{i}}{}^{\mathrm{citer}} + \lambda_{\mathrm{i'}}{}^{\mathrm{citer}} + \mu_{\mathrm{k}}{}^{\mathrm{citer}})] * \Gamma \ge (\lambda_{\mathrm{i}}{}^{\mathrm{citer}} + \mu_{\mathrm{k}}{}^{\mathrm{citer}}) * \mathrm{D(i,k)} - (\lambda_{\mathrm{i'}}{}^{\mathrm{citer}} + \mu_{\mathrm{k}}{}^{\mathrm{citer}}) * \mathrm{D(i',k)} \\ &\sum_{\mathrm{Q(iter)}} \mu_{\mathrm{k}}{}^{\mathrm{citer}} - \sum_{\mathrm{P(iter)}} \mu_{\mathrm{k}}{}^{\mathrm{citer}} < \sum_{\mathrm{k}} \mu_{\mathrm{k}}{}^{\mathrm{citer}} \quad \forall \mathrm{iter} < \mathrm{citer} \\ &\mathrm{P_{iter}} = \{\mathrm{k} \mid \mu_{\mathrm{k}}{}^{\mathrm{iter}} = 1\} \\ &\mathrm{Q_{iter}} = \{\mathrm{k} \mid \mu_{\mathrm{k}}{}^{\mathrm{iter}} = 0\} \\ &\text{D(i,k)= symbolic representaion of gene "i" in condition "k"} \\ &\lambda_{\mathrm{i}}{}^{\mathrm{citer}} = \begin{cases} \text{1, if gene i belongs to bicluster "citer"} \\ \text{0, otherwise} \end{cases} \\ &\mathrm{P_{iter}}, \mathrm{Q_{iter}} = \text{denote the set of conditions that comprised previous biclusters} \end{aligned}$$

(5)

Generation of Directed Graph

The output of the bi-clustering yields a bi-partite network. It is possible to generalize the bi-partite network into a directed graph in which the interaction between the soluble factors and the transcription factors were linked. Normally the generalization of a bi-partite network into a directed graph consists of having nodes in the input layer also being present in the output layer. This is not the case with this specific example. However, utilizing the *a priori* information given in **Table 1**, both layers can still be merged into a cohesive graph. The important assumption which is made is that each soluble factor can only stimulate its reporter. These direct links are a consequence of the experimental design. Therefore while the addition of TNF-α has been shown to also activate the STAT3 reporter, it cannot do it directly. From this, it is possible to infer the presence of links between the different nodes.

In the example provided, the justification for the link is that since the STAT3 reporter can only be activated by its associated reporter, IL-6, any activation via TNF-α must occur via IL-6. Furthermore, since TNF-α is known to activate NFkB directly, we can infer that the activation of NFkB causes in some indirect manner the activation of STAT3 either through an intermediate such as IL6. Converting each of the links found in the bi-partite solution in such a fashion, will allow for the formulation of the more commonly shown directed graph. Such graphs can then be used for the reconstruction of network dynamics.

Reconstruction of Network Dynamics

After the network architecture has been identified, it then becomes possible to solve for the underlying network dynamics, i.e. the strength of the interactions between the different transcription factors. To do so, the generalized model x' = **A**x+B, where x and x' represent the dynamics and the first derivative of the responses obtained via the LCA, and A represents the strength of the connections which vary with respect to time, and B represents a forcing function which is the addition of the soluble factors to the system. Allowing the entries A to vary with time allow us to identify significant nonlinearities within the system. This is important because oftentimes the nonlinearities point to significant mechanisms at work such tolerance. If **A** is allowed to vary with time, it is necessary for the dynamics denoted via x to be obtained under as many different conditions as there are transcription factors. The LCA with its ability to obtain dynamics for multiple transcription factors under multiple conditions satisfies this constraint.

$$\begin{aligned} &\min : \sum_{i=1}^{ng} \sum_{j=1}^{nc} \sum_{t=1}^{nt} \epsilon^{+}(i,j,t) + \epsilon^{-}(i,j,t) \\ &s.t. \\ &D'(i,j,t) \text{-} \left[\sum_{k=1}^{ng} A(i,i',t) * D(i',j,t) + \beta(j) * s(i,j) \right] \text{-} \epsilon^{+}(i,j,t) + \epsilon^{-}(i,j,t) = 0 \\ &A(i,i',t) \le \Gamma * C(i,i') \\ &A(i,i',t) \ge \text{-}\Gamma * C(i,i') \end{aligned} \tag{6}$$

The optimization formulation attempts to compute the weight of A for every time point. It is hypothesized that the numerical evolution of A over time may give insights as to the underlying mechanism or nonlinear formulation. The formulation for conducting such a reconstruction is given in (6), where D represents the data, D' represents the derivative, i is the index for the different transcription factors, j is the index for the different conditions, C is the connectivity structure determined via the bi-clustering formulation, s represents a binary matrix to denote which transcription factor is directly stimulated under a given experimental condition, and β represents a weighting function that takes into account that not all of the reporters are activated at the same level by an equivalent level of their soluble factor.

Results and Conclusion

The networks which result from the bi-clustering are given in **Figure 2,** which when converted into a directed graph yields **Figure 3.** Utilizing this network structure, it then becomes possible to solve for the different interactions.

From **Figure 3**, we begin to see significant feed forward interactions such as with those involving NFkB and STAT3, and the central location of ISRE within the network. From the data, the nodes corresponding to AP1 were not included because they were not found in any of the bi-clusters. HSE was also removed because it did not have an explicit activator, nor an experimental condition which was designed to activate it, it was impossible to infer what the effect of HSE activation upon the rest of the network is given a bi-clustering formulism.

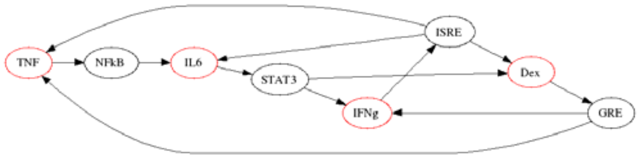

Figure 3: The Directed Graph representation of the bi-partite network in Figure 2. This network architecture is used in (6) to solve for the dynamics

Utilizing this network architecture and solving for the dynamics yields **Figure 4.** The dynamics present yield some interesting insights as to the overall mechanisms that work in conjunction with the architecture. For instance, we see that the response of NFkB to external stimulation appears to have a significant lag event perhaps due to a rate limiting dimerization event, the loss of GRE activity over time points to a tolerance mechanism coupled with the clear down-regulation of NFkB by GRE, and possible oscillatory effects associated with ISRE due to its central role in the feedback loop. One of the most interesting of these dynamics is that most of the transcription factors appear to exhibit a significant level of tolerance under constant stimulation.

Though the system which was solved is a small proof of principle example, many significant mechanisms were still evident in the final solution denoted in **Figure 4.** The fact that such dynamics were visible even in such a small case suggest that the same framework would yield useful insights in a more comprehensive system in which all of the interacting transcription factors were measured. All the more exciting is the fact that along with the computational framework, the experimental framework allows this to be accomplished efficiently and at low cost.

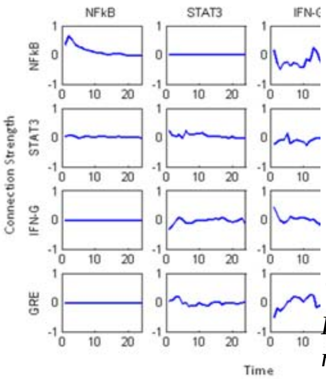

Figure 4: The time varying connection strengths between the different transcription factors

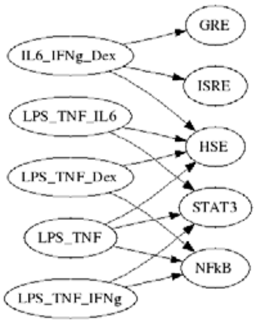

Figure 2: Bi-partite representation of the bi-clustering solution

While most of the current work in bi-clustering has focused upon creating heuristics to accurately approximate the results of an NP-Hard problem(Cheng and Church 2000; Liu and Wang 2007), such that the runtimes are reasonable, it is our contention that more attention needs to be paid to the problem of intersecting bi-clusters(Prelic, Bleuler et al. 2006), specifically the fact that neither the architecture, nor the number of bi-clusters is known *a priori*, and therefore should not be parameters within the system

By utilizing a framework which allows for overlapping bi-clusters, we were able to see dynamic signatures which correspond to mechanistic actions such as tolerance and time-lag event. We hypothesize that the dynamics of other transcription factors which were not easily interpretable may be due to interactions with transcription factors which we did not measure.

Acknowledgments

EY and IPA acknowledge support from NSF grant 0519563 and the EPA grant GAD R 832721-010. MLY acknowledges support from the NIH grants AI 063795 and EB 002503.

References

Cano, C., L. Adarve, et al. (2007). "Possibilistic approach for biclustering microarray data." Comput Biol Med **37**(10): 1426-36.

Cheng, Y. and G. M. Church (2000). "Biclustering of expression data." Proc Int Conf Intell Syst Mol Biol **8**: 93-103.

Freeman, T. C., L. Goldovsky, et al. (2007). "Construction, visualisation, and clustering of transcription networks from microarray expression data." PLoS Comput Biol **3**(10): 2032-42.

King, K. R., S. Wang, et al. (2007). "A high-throughput microfluidic real-time gene expression living cell array." Lab Chip **7**(1): 77-85.

Liu, X. and L. Wang (2007). "Computing the maximum similarity bi-clusters of gene expression data." Bioinformatics **23**(1): 50-6.

Madeira, S. C. and A. L. Oliveira (2004). "Biclustering algorithms for biological data analysis: a survey." IEEE/ACM Trans Comput Biol Bioinform **1**(1): 24-45.

Matys, V., E. Fricke, et al. (2003). "TRANSFAC: transcriptional regulation, from patterns to profiles." Nucleic Acids Res **31**(1): 374-8.

Prelic, A., S. Bleuler, et al. (2006). "A systematic comparison and evaluation of biclustering methods for gene expression data." Bioinformatics **22**(9): 1122-9.

Zhang, D. J. a. A. (2002). "Cluster Analysis for Gene Expression Data: A Survey." Jiang, D. X., and Zhang, A. Cluster Analysis for Gene Expression Data: A Survey. Technical Report 2002-06, State University of New York at Buffalo, 2002.