A MIXED INTEGER OPTIMIZATION ALGORITHM TO REVERSE ENGINEER TRANSCRIPTIONAL REGULATORY NETWORKS

P.T.Foteinou[1], E.Yang[1], G.K.Saharidis[2], M.G. Ierapetritou[2] and I.P. Androulakis[1,2*]
[1]Department of Biomedical Engineering, Rutgers University, NJ 08854
[2]Department of Chemical and Biochemical Engineering, Rutgers University, NJ 08854
[*]Corresponding Author

Abstract

A major goal in post-genomic era is to reverse engineer transcriptional regulatory networks. Advances in high-throughout technologies e.g. DNA microarrays coupled with the working knowledge on the connectivity interactions between putative regulators and their target genes have opened new opportunities for the application of reverse-engineering approaches in modeling regulatory networks. In this paper, we propose a mixed–integer optimization algorithm that combines prior biological information and expression data to identify multiple regulatory structures which can be subsequently analyzed in order to identify robust transcription factor activity profiles, as well as alternative regulatory networks that could be used for developing and exploring relevant hypotheses.

Keywords

(mixed-integer optimization, gene regulation)

Introduction

The principal goal of reverse engineering approaches is to identify the activation program of transcription modules under particular conditions (Wang, Cherry et al. 2002) so as to hypothesize how activation/deactivation of gene expression can be induced/suppressed (Ng, Bursteinas et al. 2006). The identification of transcription factors, their target genes and the interactions that control (regulate) gene expression has been addressed over the past few years with a variety of experimental and computational approaches (Iyer, Horak et al. 2001; van Steensel, Delrow et al. 2003). Recently, methods combining TF-gene connectivity data and gene expression measurements have emerged (Bussemaker, Li et al. 2001; Yeung, Tegner et al. 2002; Alter and Golub 2004; Gao, Foat et al. 2004; Kato, Hata et al. 2004; Boulesteix and Strimmer 2005; Kao, Tran et al. 2005; Tran, Brynildsen et al. 2005; Sun, Carroll et al. 2006).

In the present study we explore an optimization-based model that identifies optimal reconstruction and network architectures in a rigorous manner. The proposed algorithm captures alternative regulatory architectures as well as it assesses robustness of specific transcription factors based on a consistency metric. We further evaluate the biological implications of the multiple alternative structures in their biological context and demonstrate how a systematic framework can define the basis for a consistent hypothesis generation mechanism related to putative regulatory interactions. Another key aspect of our model is that we can take known directionality in regulation of a transcription factor into account. Complementary to this we can also infer the role for those regulators that their activity on certain promoter regions is unknown – it can be either activation or repression (unknown). Identifying robust transcription factors is of

clinical relevance as it can serve as a diagnostic tool for *in silico* target identification (Sun, Carroll et al. 2006).

Methods

The dynamics of gene expression are modeled using simple synthesis and degradation terms expressed by a set of reactions which involve the specific binding of TFs to DNA sequences as well as the recruitment of RNA polymerase I complex (Sun, Carroll et al. 2006) . Assuming a quasi-steady state for mRNA synthesis and degradation and performing a log transformation we get a log-linear model as shown in Eq. 1.

$$E = \Pi \cdot P,\ E = \log\left[\frac{[mRNA(i,t)]}{[mRNA(i,0)]}\right],\ P = \log\left[\frac{TFA(j,t)}{TFA(j,0)}\right],\ \Pi = \{\pi_{ij}\} \quad (1)$$

where E matrix is the log-ratio of the gene expression level of gene i at time point t relative to the initial condition (t=0), and its dimensions are N_g (number of genes) x N_T (number of time points), Π is the connectivity matrix whose entries are constant and characterize the strength of interaction between any regulatory pair (i,j) with j refers to the regulator and its dimensionality is N_g x N_{TF} (number of transcription factors) considering the strength coefficients as surrogates for the binding affinity of the transcription factor to the promoter region. The P matrix refers to the inferred activity profiles (TFA) for each TF expressed also as log ratios with respect to control time point (t = 0hr).

Integer Optimization

The optimization algorithm aims at decomposing the available gene expression signatures in a reduced "basis set" defined by transcription factor activities. In essence, based on the integer linear problem as shown in (Table 1) we are addressing the following questions: *Can we identify specific regulators with robust reconstructed activities across multiple network architectures? Are there preferential patterns emerge in terms of TFs functionality (activator or repressor)?*

Model Linearization

The introduction of P^{eff} variable in Table 1 introduces a non-convex bilinearity in the formulation due to the product of the continuous variable P(j,t) and the binary variable r(i,j) as shown in Eq. (2). Such a variable serves as the effective activity of a regulator on its target genes taking the nature of interaction (activator or repressor) into account. That is to say, based on the nature of regulatory interaction the effect of changes in TFA of a regulator would have distinct effects on changes on the expression level of its target genes.

$$r(i,j) = \begin{cases} 1 & \text{TF(j) activates gene(i)} \\ 0 & \text{otherwise} \end{cases}$$
$$P^{eff}(i,j,t) = [2 \cdot r(i,j) - 1] \cdot P(j,t) \quad (2)$$

Such product is exactly linearized through the introduction of the constraints (Glover 1975) and they are presented in Eq. (3)

$$-r(i,j) \cdot M - P(j,t) \le P^{eff}(i,j,t) \le r(i,j) \cdot M - P(j,t)$$
$$[r(i,j) - 1] \cdot M + P(j,t) \le P^{eff}(i,j,t) \le [1 - r(i,j)] \cdot M + P(j,t) \quad (3)$$

where M is a big number and the general form in (3) can be reduced to further sub-forms based on whether j regulator activates or represses gene i. Moreover, the superstructure of all the possible regulatory interactions is defined in Eq. (4):

$$D(i,j) = \begin{cases} 1 & \text{TF(j) regulates gene(i), i.e. } \pi(i,j) \neq 0 \\ 0 & \text{otherwise, i.e. } \pi(i,j) = 0 \end{cases} \quad (4)$$

Finally, we approximate the log-ratio of the expression data as it follows:

$$E(i,t) = \sum_j \pi_{ij} \cdot P^{eff}(i,j,t) + \text{error} \quad (5)$$

In Equation (5) the presence of "error" term simulates the existence of potential sources of uncertainty associated with lack of knowledge about structure connectivity and directionality.

Table 1: Mixed-Integer Formulation

mixed-integer Synthesis & Analysis of Regulatory Networks (miSARN)

$$\min \sum_i \sum_t e^+(i,t) + e^-(i,t)$$

subject to

$$E(i,t) - \sum_j \pi(i,j) P^{eff}(i,j,t) = e^+(i,t) - e^-(i,t) \quad \forall i,t$$
$$\sum_j z(j) = m \le N_{TF}$$
$$\sum_j D(i,j) \cdot z(j) \ge 1 \quad \forall i$$
$$-r(i,j)M - P(j,t) \le P^{eff}(i,j,t) \le r(i,j)M - P(j,t) \quad \forall i,j,t$$
$$[r(i,j) - 1]M + P(j,t) \le P^{eff}(i,j,t) \le [1 - r(i,j)]M + P(j,t) \quad \forall i,j,t$$
$$z(j) P_{max} \le P(j,t) \le z(j) P_{min} \quad \forall j,t$$
$$\sum_{j \in N^k} z(j) - \sum_{j \in B^k} z(j) \le |N^k| - 1$$
$$N^k = \{j \mid z^k(j) = 1\},\ B^k = \{j, \mid z^k(j) = 0\}$$
$$D(i,j) = \begin{cases} 1 & \pi(i,j) \neq 0 \\ 0 & \pi(i,j) = 0 \end{cases} \quad \forall i,j$$
$$P(j,t), P^{eff}(i,j,t) \in \Re$$
$$e^+(i,t), e^-(i,t) \in \Re^+ \quad \forall i,j,t$$
$$z(j), r(i,j) \in \{0,1\} \quad \forall i,j$$
$$i = 1,\ldots,N_g; j = 1\ldots,N_{TF}; t = 1\ldots N_T$$

The mixed-integer linear optimization problem (Table 1) solved using the GAMS modeling software (Brooke, Kendrick et al. 2004) running CPLEX for the solution of the corresponding MILP. We used temporal expression data of *E coli* during transition from glucose to acetate as the sole carbon source publicly available at http://ww.seas.ucla.edu/~liaoj/. The connectivity matrix was based on RegulonDB database (Salgado, Santos-Zavaleta et al. 2001).

Results-Discussion

The complete regulatory network consists of 30 regulators and given the fact that each gene must be regulated by at least one TF the formulation becomes infeasible if parameter m (Table 1) becomes less than 18. That is to say, we run the optimization algorithm parametrically with respect to number of TFs getting the pattern as shown in Figure 1. Interestingly, we observe that for m=26…30 the reconstruction error remains the same.

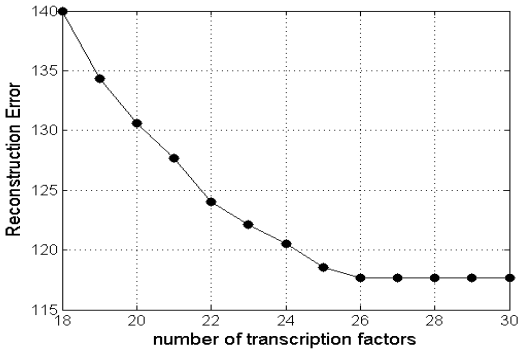

Figure 1: Reconstruction error vs. number of transcription factors

We are identifying robust activity profiles (TFA) across 13 multiple regulatory structures (m=18…30) applying a consistency metric for each j regulator; robustness-R(j) defined as it follows: R(j)=[f(j)/Mt]*C(j), where f(j) denotes the frequency of j TF across Mt total multiple solutions and C(j) defines the average Pearson's correlation coefficient for the multiple inferred TFA - P(j,t). The reconstructed profiles for the 13 multiple solutions are shown in Figure 2 distinguishing 9 TFs (*Ada, CysB, FadR, GatR, LeuO, Lrp, PurR, TrpR, and TyrR*) to be characterized by the highest robustness (R(j)=1). This critical subset of TFs indeed play a critical role in the transition of *E coli* from glucose to acetate (Landini, Hajec et al. 1994). Moreover, we further generate alternative structures with the same reconstruction error for a given m by activating the integer cuts. By doing so, we get equivalent network structures (Figure 3) that open the possibility of identifying TFs whose targeted silencing might be "lethal" to system's dynamics. For instance, the Phage-Shock-Protein System shown in the upper left (Figure 3) is regulated by PspF and RpoN promoters in Y. Enterocolitica, a bacteria very similar to E. Coli; it was found that a PspF null mutation did not impart lethality upon the specific strain, but rather caused a slight decrease in the growth rate of the strain, as was the deletion of the RpoN promoter region. In fact the deletion of either the PspF or the RpoN sequence from the promoter region yielded a strain that was nearly indistinguishable(Maxson and Darwin 2006), suggesting that with the deletion of a single promoter sequence, in the pspA gene, the other transcription factor can indeed compensate for the loss in control.

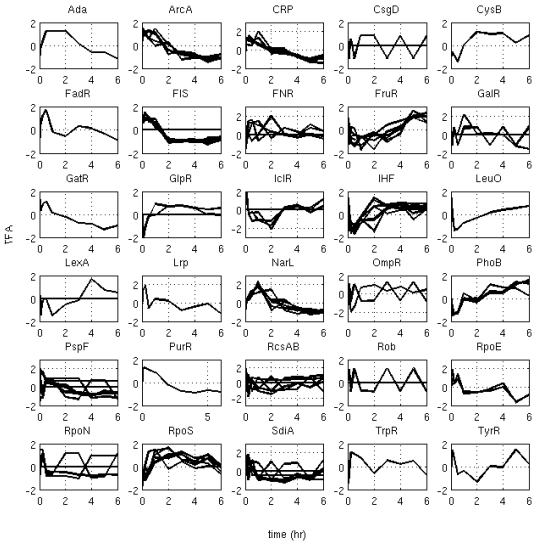

Figure 2: Reconstruction of TFA profiles

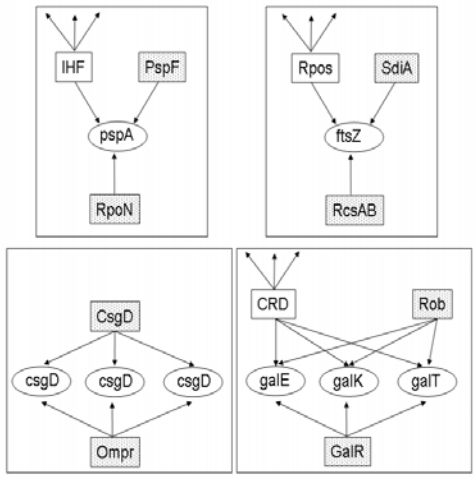

Figure 3: Equivalent network architectures. Square:TF, oval: Genes, Dark squares: interchangeable TFs, CsgD denotes the activity of the corresponding TF, csgD denotes the gene

As far as the TFs with unknown directionality is concerned we get the following 3 TFs along with their target genes: (1) *CRP*: galE, galK, galT, prop; (2) *LRP*: kbl; (Stelling, Sauer et al.) *PhoB*: ugpB, ugpE. The inferred role of the 3 aforementioned TFs across the 13 multiple solutions is consistent for all TFs**.** Specifically, there is only one solution out of 13 in which the transcription factor CRP acts as a repressor. The remaining

solutions identify the following relations: (1) *CRP*: activates galE, galK, galT; represses proP; (2) *LRP*: activates kbl; (3) *PhoB*: represses ugpB, ugpE.

Conclusions

Our proposed optimization algorithm allows us to unravel the principles that govern complex biological phenomena such as gene regulation. We suggest a systematic framework that integrates high-throughput data, network connectivity coupled with the known directionality of regulation for most of the regulatory pairs with the fundamental task to gain biological insight about regulatory networks. Therefore, our model not only provides us with the optimal reconstruction but also it offers us the possibility of generating multiple architectures that ultimately reveal us either a critical subset of TFs with robust activity profiles or a set of interchangeable TFs crucial to which TFs have a "lethal" impact to the dynamics of the system. Finally, our model can decipher the directionality of those TFS whose regulatory role is unknown.

Acknowledgments

The authors acknowledge financial support from NSF under the NSF-BES 0519563 Metabolic Engineering Grant and the EPA under the EPA GAD R 832721-010.

References

Brooke, A., Kendrick, D., Meeraus, A. A. (1992). GAMS- A User's Guide (Release 2.25). *The Scientific Press*. San Francisco, CA.

Alter, O. and G. H. Golub (2004). "Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription." Proc Natl Acad Sci U S A **101**(47): 16577-82.

Boulesteix, A. L. and K. Strimmer (2005). "Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach." Theor Biol Med Model **2**: 23.

Brooke, A., D. Kendrick, et al. (2004). GAMS A user's guide, GAMS Development Corporation.

Bussemaker, H. J., H. Li, et al. (2001). "Regulatory element detection using correlation with expression." Nat Genet **27**(2): 167-71.

Gao, F., B. C. Foat, et al. (2004). "Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data." BMC Bioinformatics **5**: 31.

Glover, F. (1975). "Improved Linear Integer Programming Formulations of nonlinear integer problems." Management Science **22**(4).

Iyer, V. R., C. E. Horak, et al. (2001). "Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF." Nature **409**(6819): 533-8.

Kao, K. C., L. M. Tran, et al. (2005). "A global regulatory role of gluconeogenic genes in Escherichia coli revealed by transcriptome network analysis." J Biol Chem **280**(43): 36079-87.

Kato, M., N. Hata, et al. (2004). "Identifying combinatorial regulation of transcription factors and binding motifs." Genome Biol **5**(8): R56.

Landini, P., L. I. Hajec, et al. (1994). "Structure and transcriptional regulation of the Escherichia coli adaptive response gene aidB." J Bacteriol **176**(21): 6583-9.

Maxson, M. E. and A. J. Darwin (2006). "Multiple promoters control expression of the Yersinia enterocolitica phage-shock-protein A (pspA) operon." Microbiology **152**(Pt 4): 1001-10.

Ng, A., B. Bursteinas, et al. (2006). "pSTIING: a 'systems' approach towards integrating signalling pathways, interaction and transcriptional regulatory networks in inflammation and cancer." Nucleic Acids Res **34**(Database issue): D527-34.

Salgado, H., A. Santos-Zavaleta, et al. (2001). "RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12." Nucleic Acids Res **29**(1): 72-4.

Stelling, J., U. Sauer, et al. (2004). "Robustness of cellular functions." Cell **118**(6): 675-85.

Sun, N., R. J. Carroll, et al. (2006). "Bayesian error analysis model for reconstructing transcriptional regulatory networks." Proc Natl Acad Sci U S A **103**(21): 7988-93.

Tran, L. M., M. P. Brynildsen, et al. (2005). "gNCA: A framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation." Metabolic Engineering **7**(2): 128-141.

van Steensel, B., J. Delrow, et al. (2003). "Genomewide analysis of Drosophila GAGA factor target genes reveals context-dependent DNA binding." Proc Natl Acad Sci U S A **100**(5): 2580-5.

Wang, W., J. M. Cherry, et al. (2002). "A systematic approach to reconstructing transcription networks in Saccharomycescerevisiae." Proc Natl Acad Sci U S A **99**(26): 16893-8.

Yeung, M. K., J. Tegner, et al. (2002). "Reverse engineering gene networks using singular value decomposition and robust regression." Proc Natl Acad Sci U S A **99**(9): 6163-8.