16th European Symposium on Computer Aided Process Engineering
and 9th International Symposium on Process Systems Engineering
W. Marquardt, C. Pantelides (Editors)

Reconstruction of Transcriptional Regulatory Networks via Integer Linear Programming

João M. S. Natali, José M. Pinto

Polytechnic University, 6 Metrotech Center, Brooklyn, NY 11201, USA

Abstract

Much effort has been recently dedicated to the identification of genome-wide transcription regulatory networks by means of comprehensive high-throughput experiments that are capable of capturing the systemic behavior of the transcription coordination phenomenon. The present work comprises the development of Linear Programming (LP) and Integer Linear Programming (ILP) approaches to model and analyze the gene regulatory network of *Saccharomyces cerevisiae*, centered on a logic inference based representation of regulatory events and on a direct evaluation of experimental quantitative results. Our models are based in a simple representation of regulatory logic. Initial results show coherence to published data and improvements on the logical representation of regulatory events are currently under development.

Keywords: Integer Linear Programming, Transcriptional Regulatory Networks, Optimization, Logical Modeling.

1. Introduction

The concept that proteic molecules are produced via the initial transcription of the genetic code into mRNA strands and the following translation of this mRNA into a sequence of amino acids – which is often referred to as the Central Dogma of Biology – has been widely known and accepted by the scientific community for decades. However, the mechanisms underlying the regulation of these processes, which would ultimately explain why proteins are produced in such differing quantities under diverse metabolic conditions, are far from being fully understood.

The transcription of a gene relies, among many factors, on the activities of a class of enzymes called RNA-polymerases. The binding of the RNA polymerase to the genetic code may depend on the existence of other chromatin binding proteins and complexes, known as transcription factors (or simply TF's), which can aid or obstruct the enzyme's binding and further genetic transcription. Therefore, the affinities and activities of transcription factors are key elements of the cell's transcription regulation.

In recent years significant effort has been put into deciphering transcription regulatory elements and regulatory networks on a genomic scale. This attempt has been founded on the insight that the information that can be extracted from the establishment of a coordinated network of regulatory interactions may reach far broader scopes of understanding than the usual recognition of individual regulatory elements alone.

One of the most successful experiments aiming at the identification of a genome-wide regulatory network was carried out by Richard Young's group (Lee et al., 2002) based upon the ideal eukaryotic microorganism *Saccharomyces cerevisiae*. These authors relied on Chromatin-Immunoprecipitation (ChIP) and Microarrays techniques to identify all gene promoter regions that were physically bounded by a comprehensive set of transcription factors. The outcome from this approach was an array of *p-values* that provided information on the likelihood of each of the promoter regions from the whole

studied genome to be bound (and, thus, potentially regulated) by each of the transcription factors considered.

Additionally, a number of authors undertook the effort of defining mathematical methods and computational procedures that would be capable of providing information on gene regulatory networks in a *in silico* manner. Gupta et al. (2005) used linear and non-linear dynamical models of mRNA production and gene expression to obtain regulatory patterns from a set of expression profiles. Hasty et al. (2001) provided a comprehensive review on computational studies of gene regulatory networks. Using a procedure based on statistical analysis and relying on the results from Lee et al. (2002) and on the vast amount of expression data currently available, Gao et al. (2003) proposed the confrontation of the information extracted from transcription factor occupancy data and gene expression data to obtain a compendious set of TF-Gene interactions that represent a concise and coherent regulatory network. For that purpose, however, a number of simplifications based on statistical calculations were performed. These simplifications are justifiably expected not to exert great influence on the general outcome of the method; however, they significantly limit the reproducibility of the results while inserting information that is not exclusively provenient of the biological phenomenon studied.

2. Proposed Approaches

We propose two approaches for the Regulatory Network Reconstruction problem, one based on Linear Programming (LP) and another on Integer Linear Programming (ILP). The formulated problems involve the modeling of regulatory events and the automated decision making regarding which of the interactions between transcription factors and intergenic regions, previously pointed by genomic location analysis (Lee et al., 2002), are indeed relevant to the global regulation of transcription in yeast and, therefore, belong to its regulatory network.

2.1. Linear Programming/Minimum Cost Network Flow Model

The proposed LP formulation is a Minimum Cost Network Flow (MCNF) model in which supply nodes are set to represent transcription factors, demand nodes are regarded as genes, and arcs represent the paths through which regulatory signals flow. The model is based on the representation of regulatory elements and signalling pathways as a network comprised of a bipartite graph and input/output flows. A graphical representation of such network is presented in figure 1.

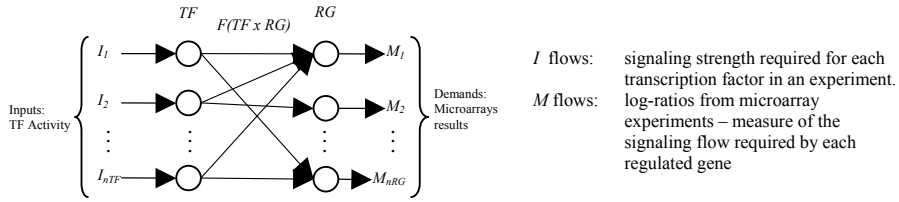


Figure 1: Regulatory interaction network in LP model..

The model is defined in the following manner: let $TF = \{1, 2, ..., n_{TF}\}$ be the set of transcription factors, and $RG = \{1, 2, ..., n_{RG}\}$ the assumed set of regulated genes. The bipartite graph that represents the interconnections between transcription factors and

genes is given by $F=(V_F,E_F)$, where $V_F = TF \cup RG$ and $E_F = TF \times RG$. The network model is based on the straightforward concept of flow balance around each defined node, in which microarray experiment results are used as a measure of the signal intensity of positive regulation required by each gene and, thus, determine the overall intensity and units of the global flow through the network. This signal flow is distributed to every gene from each transcription factor through the network *F*. A cost parameter $C_{i,j} \ \forall (i,j) \in (TF, RG)$ is further associated with each unitary flow through arcs in *F*, and is defined as the *p-value* for the existence of an interaction found by Lee at al. (2002). Finally, a demand for flux $M_j \ \forall j \in RG$ is assigned to each gene in the network, given by the base-2 logarithm of the ratio of scanned luminescence intensity between the test and control media in each run of the microarrays experiments.

The problem is, then, formulated as a MCNF problem (Ahuja et al., 1993). Only the data from a single microarray experiment is considered in the definition of the problem. Therefore, the comparison between different solutions using dissimilar expression data is important for the interpretation of the results. The resulting optimization problem is as follows:

$$\begin{aligned} &\min \ Z = \sum_{i \in TF} \sum_{j \in RG} C_{i,j} \cdot F_{i,j} \\ &s.t. \ \left\{ F_{i,j} \ \forall i \in TF, \forall j \in RG \ \middle| \ \sum_{i \in TF} F_{i,j} \geq M_j; \ \sum_{j \in RG} F_{i,j} \geq 0; \ 0 \leq F_{i,j} \leq F^{UP} \right\} \end{aligned} \tag{1}$$

2.2. Integer Linear Programming/Logic Inference Based Representation

Let $EX = \{1, 2, ..., n_{EX}\}$ be the set of experiments used as input data, and *TF* and *RG* be defined as previously. Furthermore, we define $X_{j,k} \in \{True, False\} \ \forall j \in RG, \forall k \in EX$ as a Boolean variable which is true if and only if gene *j* is expressed in experiment *k*, and $Y_{i,k} \ \forall i \in TF, \forall k \in EX$ as a Boolean variable true if and only if a transcription factor *i* is produced in an experiment *k*. Moreover, we define two sets of binary variables representing the topological characteristics of the reconstructed transcriptional regulatory network. Let $Sp_{i,j} \ \forall (i,j) \in (TF, RG)$ be a set of Boolean variables which are true if and only if the transcription factor *i* activates the transcription of gene *j*, and, concordantly, $Sn_{i,j} \ \forall (i,j) \in (TF, RG)$ which are true if and only if the transcription factor *i* represses the transcription of gene *j*.

Using this formalism, the logical relationship between the variables that represent the regulatory network topology and the inferences that connect TF's and genes can be posed as follows:

$$\begin{bmatrix} Sp_{i,j} \\ L_A(X_{j,k}, Y_{i,k}) \ \forall k \in EX \end{bmatrix} \vee \begin{bmatrix} Sn_{i,j} \\ L_R(X_{j,k}, Y_{i,k}) \ \forall k \in EX \end{bmatrix} \vee \begin{bmatrix} \neg Sp_{i,j} \wedge \neg Sn_{i,j} \\ \text{No Regulation} \end{bmatrix} \ \forall (i,j) \in (TF, RG) \tag{2}$$

In disjunction (2), $L_A(X_{j,k}, Y_{i,k})$ represents a set of logical propositions that describe relationships between the expression of a gene and the binary activity of a transcription factor under activation interactions, and $L_R(X_{j,k}, Y_{i,k})$, similarly, for repression.

The present model is based on the simple logical relationships between the activities of transcription factors and the genes that are regulated by them as below:

$$\begin{bmatrix} Sp_{i,j} \\ X_{j,k} \Rightarrow Y_{i,k} \end{bmatrix} and \begin{bmatrix} Sn_{i,j} \\ X_{j,k} \Rightarrow \neg Y_{i,k} \end{bmatrix} \quad \forall (i,j,k) \in (TF, RG, EX) \tag{3}$$

which can be converted to integer constraints (Raman and Grossmann, 1991). The OF is the maximization of the existence of activation interactions, weighted by the log ratio values of each interactions shown below, where R is the set of log ratio values for the interactions between each pair of transcription factors and genes found by Lee et al. (2002). The optimization model can be defined by:

$$\max \quad Z = \sum_{i \in TF} \sum_{j \in RG} R_{i,j} \left(Sp_{i,j} + Sn_{i,j} \right) \tag{4}$$

$$s.t. \quad Y_{i,k} - X_{j,k} \geq \left(1 - Sp_{i,j} \right) \qquad \forall i \in TF, \forall j \in RG, k \in EX \tag{5}$$

$$Y_{i,k} + X_{j,k} - 1 \leq \left(1 - Sn_{i,j} \right) \qquad \forall i \in TF, \forall j \in RG, k \in EX \tag{6}$$

$$\sum_{j \in RG} \sum_{k \in EX} M_{j,k} X_{j,k} \geq MX^{Lo} \tag{7}$$

$$Sp_{i,j} + Sn_{i,j} \leq 1 \qquad \forall i \in TF, \forall j \in RG \tag{8}$$

$$X_{j,k}, Y_{i,k}, Sp_{i,j}, Sn_{i,j} \in \{0,1\} \qquad \forall i \in TF, \forall j \in RG, k \in EX \tag{9}$$

The above formulation results in a relaxed problem, provided that no constraints on Y are defined. To address this issue, and considering that the activation and repression events are only connected by the objective function and by the mutually exclusiveness constraint, the problem was divided into two subproblems shown below.

$$\begin{array}{l|l} \max \ Z = \sum_{i \in TF} \sum_{j \in RG} R_{i,j} \left(Sp_{i,j} \right) & \max \ Z = \sum_{i \in TF} \sum_{j \in RG} R_{i,j} \left(Sn_{i,j} \right) \\ s.t. \ (5), (7), (9) & s.t. \ (6), (7), (9) \\ Y_{i,k} = 0 \quad \forall i \in TF, k \in EX & Y_{i,k} = 1 \quad \forall i \in TF, k \in EX \end{array} \tag{10}$$

The *a priori* definition of the transcription factors activities is based on the logical nature of the model. Considering the activation problem, the postulation of no TF activity implies that all genes which are positively regulated be non-expressed. This generates a trade-off with the OF, which seeks the maximization of regulatory interactions and the imposition of a lower bound on genetic expression. The same argument is used to justify the repression model. It is important to note, however, that using the simple logic proposed, alongside the restrictions on the values of Y, both models are reduced to the same set of constraints and, thus, to the same formulation.

3. Results and Discussion

3.1. Case Study: Yeast Transcription Network

The system employed in this study is the transcription network from *S. cerevisiae*, comprised of the microorganism's entire genome (6270 genes) and a set of 113 transcription factors, chosen due to the availability of the chromatin binding data obtained by Lee et al. (2002).

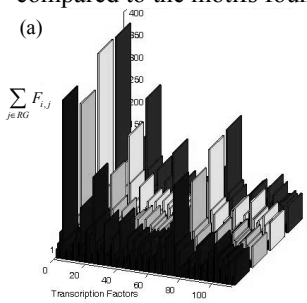
Genome-wide binding data consist of two-dimensional arrays containing all the base 2 log ratios of the scanned intensities of colored tags from the ChIP concentrated solution and the control solutions for all considered transcription factors. Microarray experiment results were obtained from the *Saccharomyces* Genome Database[1]. Expression profiles from four different cultivation conditions were used: evolution in limited glucose;

[1] **http://www.yeastgenome.org/**

diauxic shift; cell cycle phases S/G2/M (elutriation); and cell cycle phases M/G1 (a-factor release).

3.2. MCNF Results

Computations based on the linear model were carried for each set of expression data obtained. The optimal flow distribution from transcription factors to genes was obtained and positive interactions were considered as active arcs in the network. The input flow from each TF – equal to the sum of outbound flows in each TF node – was plotted for each experimental data. Moreover, a search for motifs was carried and the results were compared to the motifs found by Lee et al. (2002).



(b) Results

	Number of Regulatory Motifs Found					
	Autoreg	Feedfwd	Multi-Comp.	Regulator Chain	Single Input	Multi Input
Lee et al. (2002)	10	49	3	188	90	81
Exp1	2	13	4	35	42	27
Exp2	1	15	1	23	40	33
Exp3	3	21	0	29	34	30
Exp4	3	19	2	32	36	38

Figure 2: (a) Flux intensity from each transcription factor considered and for each genetic expression datasets. (b) Number of regulatory motifs found for each experimental dataset and published results from location analysis.

Figure 2 shows coherent results regarding assignment of transcription factors to genes when different microarray results are used. Moreover, the set of regulatory motifs found is seen to be smaller than the ones found at Lee et al. (2002). This result was expected provided that we are attempting to contrast such results to expression experiments to obtain a more concise set of interactions.

3.3. ILP Results

The ILP model was solved simultaneously for the four sets of experimental data selected. The results shown in figure 3 refer to the behavior of the activation model, since, as discussed, the logical inference used was not sufficient to provide a distinction between both models.

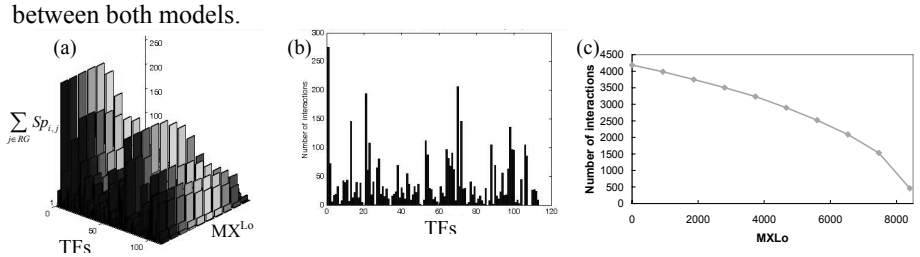


Figure 3: Total number of interactions: (a) from each transcription factor taken for different lower bounds on gene expression; (b) for each transcription factor from literature, obtained by location analysis. (c) from the entire solution space, as a function of the lower bound on expression.

Figure 3a shows that increasing the requirement for expression, the number of regulatory interactions observed suffer considerable drop. This can also be seen by the results displayed in figure 3c, which shows that the overall number of interactions also

decreases in a similar fashion. However, it can be observed that this trend is not followed with the same intensity by all the TF's, which is closely related to the greater requirement for expression by some genes relative to others, given by the log-ratio intensity in the parameter *M*. It is also observed that transcription factors that are associated with a larger number of interactions have a tendency to maintain these interactions for small decreases in the value of MX^{Lo}, whereas less connected transcription factor present a stronger dependence on this parameter. Results exhibited in table 1 corroborate the observed tendency of obtaining a reduced set of regulatory motifs, in comparison to published results from location analysis. Figure 3b illustrates that the obtained results share a good correlation with the ones found by location analysis (Lee et al., 2002).

Table 1: Regulatory motifs found in literature and those obtained from the ILP model.

Results	Number of Regulatory Motifs Found					
	Autoregulation	Feedforward Loop	Multi-Component	Regulator Chain	Single Input	Multi Input
Lee et al. (2002)	10	49	3	188	90	81
ILP (MX^{Lo} = 1000)	6	36	2	102	76	64
ILP (MX^{Lo} = 6500)	3	19	0	56	29	25

4. Conclusions

Two mathematical programming models for the reconstruction of transcriptional regulatory networks were proposed. The MCNF model represented a simple approach whose simplicity and the impossibility of incorporating multiple expression datasets limits its applicability. Results, nonetheless, show good coherence with information available in literature and a relatively high consistency with the expected behavior of the system. The second proposed model is initially formulated in disjunctive form and transformed into an Integer Linear Program. It provides a framework capable of incorporating sophisticated logical relationships between transcription factors and regulated genes that are able to describe complex regulatory relationships. The model was evaluated with a simple relational logic, which carried restrictive simplifications, and the obtained results, regardless of such complexity reductions, presented good agreement with published results and with the physics of the problem.

References

R.K. Ahuja, T.L. Magnanti and J.B. Orlin, 1993, Network Flows, Prentice-Hall, New Jersey

F. Gao, B.C. Foat and H.J. Bussemaker, 2004, Defining Transcriptional Networks Through integrative Modelling of mRNA Expression and Transcription Factor Binding Data, BMC Bioinf., Vol. 5, No. 31

A. Gupta, J.D. Varner and C.D. Maranas, 2005 , Large-Scale Inference of the Transcriptional Regulation of *Bacillus subtilis*, Comp & Chem. Eng., Vol. 29, pp. 565-576.

J. Hasty, D. McMillen, F. Isaacs and J.J. Collins, 2001, Computational Studies of Gene Regulatory Networks: In Numero Molecular Biology, Nature Reviews, Vol. 2, pp. 268-279

T.I. Lee et al., 2002, Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*, Science, Vol. 298, pp. 799-804

R. Raman and I.E. Grossmann, 1991, Relation Between MILP Modelling and Logical Inference for Chemical Process Design, Comp. and Chem. Eng., Vol. 15, No. 2, pp. 73-84