J Glob Optim (2009) 43:263–276
DOI 10.1007/s10898-007-9266-6

A mixed-integer optimization framework for the synthesis and analysis of regulatory networks

Panagiota T. Foteinou · Eric Yang · Georges K. Saharidis · Marianthi G. Ierapetritou · Ioannis P. Androulakis

Received: 3 April 2007 / Accepted: 27 October 2007 / Published online: 21 December 2007

Abstract *Motivation*: A novel mixed-integer optimization framework is proposed for the design and analysis of regulatory networks. The model combines gene expression data and prior biological knowledge regarding the potential for regulatory interactions between genes and their corresponding transcription factors. The formalism provides significant advantages over available modeling methodologies in that the complexity of the regulatory network can be explicitly taken into account, multiple alternative structures can be systematically generated and finally robust and biological significant regulators can be rigorously identified. The original non-convex mixed integer reformulation is appropriately linearized and the resulting MILP is effectively optimized using standard solvers. The versatility is demonstrated using gene expression and binding data from an *E. coli* case study during transition from glucose to acetate as the sole carbon source.

Keywords Bioinformatics · Mixed integer linear optimization · Gene regulation

1 Introduction

Significant efforts have been made experimentally and computationally, to identify transcription factors (TF), their target genes and the interaction mechanism that control (regulate) gene expression (Iyer et al. 2001; van Steensel et al. 2003). However the production of a TF is a necessary but not sufficient condition for transcription initiation and regulation. Therefore, regulator transcription levels are generally not appropriate measures of transcription factor activity (TFA). Recently, methods combining TF-gene connectivity data and gene expression measurements have emerged in order to quantify these regulatory interactions (Bussemaker et al. 2001; Yeung et al. 2002; Alter and Golub 2004; Gao et al. 2004;

P. T. Foteinou · E. Yang · I. P. Androulakis (✉)
Department of Biomedical Engineering, Rutgers University, Piscataway, NJ 08854, USA
e-mail: yannis@rci.rutgers.edu

G. K. Saharidis · M. G. Ierapetritou · I. P. Androulakis
Department of Chemical and Biochemical Engineering, Rutgers University, Piscataway, NJ 08854, USA

Kato et al. 2004; Boulesteix and Strimmer 2005; Kao et al. 2005; Tran et al. 2005; Sun et al. 2006). The main goal of this reverse engineering is to identify the activation program of transcription modules under particular conditions (Wang et al. 2002) so as to hypothesize how activation/deactivation of gene expression can be induced/suppressed (Ng et al. 2006). Aside from the development of descriptive models that correlate TFA and expression of target genes, a critical question becomes how to identify those TFs that significantly contribute to regulation and should be modulated. Along those lines (Gao et al. 2004) speculate that the mRNA profile of the target gene should be similar to the reconstructed TFA for the regulating proteins, (Sun et al. 2006) claim that accurate binding information should lead to robust TFA reconstructions whereas (Chen et al. 2005) develop a greedy-based selection of critical regulators.

In the present study we explore an optimization-based model that identifies optimal reconstruction and architectures in a rigorous manner. We propose systematic construction of alternative regulatory architectures and propose a consistency metric for assessing the robustness specific transcription factors. 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 might serve as a diagnostic tool for *in silico* target identification (Sun et al. 2006).

2 Methods

2.1 Network model

The rate of production of mRNA is modeled using simple synthesis and degradation terms (Thomas et al. 2004; Sun et al. 2006) 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. The dynamics of gene expression can thus be described as in (1) expressing a balance between promoter activity and mRNA degradation (Tran et al. 2005) modeled by a power-law kinetics (Savageau 1976). The activities of transcriptions factors (TFA) are terminal signals controlling transcriptional regulation. Therefore, TFAs represent gene expression dynamics in an attractive way as they represent a surrogate of the integrated contribution of a TF to the regulation of gene expression. Since TFs affect both the synthesis and degradation terms of the corresponding mRNA, with rate constants $\mathrm{k_s}$ and $\mathrm{k_d}$ respectively, in the general case we may assume that a different set of factors contributes to the synthesis and degradation respectively (1). The index “i” denotes a gene being regulated by transcription factor “j” and “k” respectively, “t” denotes time. $\overline{\mathrm{TFA}}$ and $\overline{\mathrm{DFA}}$ denotes the activity of the regulator and degradation factors respectively, and [mRNA] denotes the concentration of the mRNA of the corresponding gene. This power-law rate expression assumes a rate of synthesis depending on the activities of TFs whereas the degradation term is also considered proportional to the actual mRNA levels (Tran et al. 2005). The interaction strengths are denoted by π_{ij}. Making the quasi-steady state approximation for $\mathrm{mRNA(i,t)}$ and solving the corresponding algebraic equation leads to (2), accounting for an appropriate normalization with respect to the initial conditions, where without loss of generality we have collectively represented the ratio of the activities by TFA (Tran et al. 2005). Finally in (3) log-transformation results in a generalized linear expression, where the 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 $\mathrm{N_g}$ (number of genes) × N_T (number of time points), $\overline{\pi} = \{\pi_{ij}\}$ is the connectivity matrix whose entries are constant and characterize the strength of interaction between any regulatory pair (i, j) with j referring to the regulator and i to the target gene. The dimensionality of connectivity matrix is $\mathrm{N_g} \times \mathrm{N_{TF}}$ (number of transcription factors). The matrix P describes the inferred effective dynamic activities for each regulator, expressed also as log-ratios, during the time course of the experiment. In certain decomposition schemes (Tran et al. 2005) $\overline{\pi} = \{\pi_{ij}\}$ is treated as an unknown variable that must be identified. In our formulation we opted to treat the strength coefficients as surrogates for the binding affinity of the transcription factor to the promoter region. In the mathematical formulation, $\overline{\pi} = \{\pi_{ij}\}$ is similar to the Hill-Coefficient. Considering the binding of transcription factors to the promoter region, we hypothesize that the strength of the binding interactions is related to the cooperative binding interactions of the separate binding domains in the transcription factor. Therefore, the interaction coefficients will be considered to be either known from experimental studies (Lee et al. 2002; Harbison et al. 2004) or determined computationally by associating binding affinities to position weight matrices (Stormo and Fields 1998).

$$\frac{\mathrm{d[mRNA(i,t)]}}{\mathrm{dt}} = \mathrm{k_s}\prod_{\mathrm{j}}\overline{\mathrm{TFA}}(\mathrm{j,t})^{\pi_{\mathrm{ij}}} - \mathrm{k_d}\prod_{\mathrm{k}}\overline{\mathrm{DFA}}(\mathrm{k,t})^{\pi_{\mathrm{ik}}}[\mathrm{mRNA(i,t)}] \tag{1}$$

$$\begin{aligned}[\mathrm{mRNA(i,t)}] &= \frac{\mathrm{k_s}}{\mathrm{k_d}}\frac{\prod_{\mathrm{j}}\overline{\mathrm{TFA}}(\mathrm{j,t})^{\pi_{\mathrm{ij}}}}{\prod_{\mathrm{k}}\overline{\mathrm{DFA}}(\mathrm{k,t})^{\pi_{\mathrm{ik}}}} = \frac{\mathrm{k_s}}{\mathrm{k_d}}\prod_{\mathrm{j}}\mathrm{TFA}(\mathrm{j,t})^{\pi_{\mathrm{ij}}}\\ &\Rightarrow \frac{[\mathrm{mRNA(i,t)}]}{[\mathrm{mRNA(i,0)}]} = \prod_{\mathrm{j}}\left[\frac{\mathrm{TFA(j,t)}}{\mathrm{TFA(j,0)}}\right]^{\pi_{\mathrm{ij}}}\end{aligned} \tag{2}$$

$$\mathrm{E} = \prod \cdot \mathrm{P},\ \ \mathrm{E} = \log\left[\frac{[\mathrm{mRNA(i,t)}]}{[\mathrm{mRNA(i,0)}]}\right], \mathrm{P} = \log\left[\frac{\mathrm{TFA(j,t)}}{\mathrm{TFA(j,0)}}\right], \overline{\pi} = \{\pi_{\mathrm{ij}}\} \tag{3}$$

In addition to the strength of the interactions, the directionality of the activation is also critical given that transcription factors are known to exhibit multifunctional characteristics (Drazinic et al. 1996). TFs are known to act as activators, repressors or exhibit both characteristics depending on conditions. Therefore, given the effective activity of a transcription factor we need to be able to simulate its corresponding effect, whether it is activating or repressing the expression of the target genes. Assuming for simplicity that one TF regulates a single gene, then depending on the nature of the interaction the effect of changes in the TFA will have distinct effects on the changes in gene expression. If the activity of the factor increases (panel a) and if the factor activates the expression of the gene, then the corresponding expression should increase (panel b), whereas if the factor represses the expression of the gene, then the increase in activity should result in decrease in the expression of the gene (panel c). Equivalent arguments can be made for the case where the activity of the factor decreases, Fig. 1. We model the activation/repression by introducing a new variable, $\mathrm{P^{eff}(i,j,t)}$ which represents the effective TFA of a regulator for gene “i” given that the type of interaction, either repressor or activator, has been identified. The definition is done through the introduction of a binary variable, r(i, j) that takes the value of 1 if the TF(j) activates gene (i), and zero

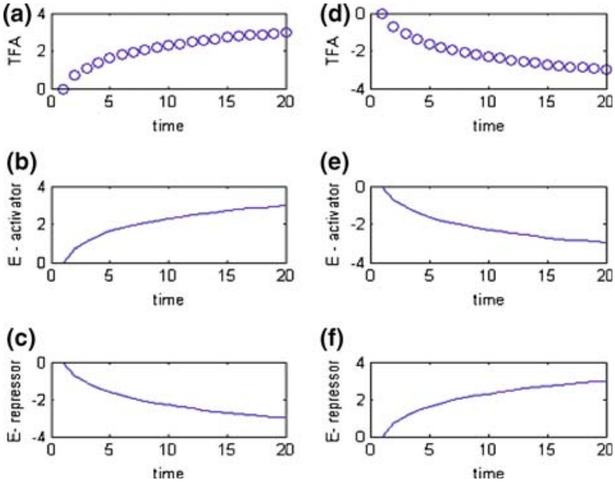

Fig. 1 Activation/repression of gene expression

otherwise (4a). The effective transcription factor activity $P^{eff}(i, j, t)$ is then defined through (4b). The superstructure of all possible regulatory interactions are defined based on (5).

$$r(i, j) = \begin{cases} 1 & \text{TF(j) activates gene(i)} \\ 0 & \text{otherwise} \end{cases} \tag{4a}$$

$$P^{eff}(i, j, t) = \left[2 \cdot r(i, j) - 1\right] \cdot P(j, t) \tag{4b}$$

$$D(i, j) = \begin{cases} 1 & \text{TF(j) regulates gene(i), i.e. } \pi(i, j) \neq 0 \\ 0 & \text{otherwise, i.e. } \pi(i, j) = 0 \end{cases} \tag{5}$$

Finally, we approximate the log-ratio of the expression data as per Eq. 6.

$$E(i, t) = \sum_j \pi_{ij} \cdot P^{eff}(i, j, t) + \text{error} \tag{6}$$

The “error” term is incorporated to simulate error-in-measurements, potential sources of uncertainty and the general lack of detailed knowledge about transcription factors, connectivity and the relationship between binding and activity.

2.2 Predicting alternative regulatory structures

It has long been hypothesized that alternative pathways connecting regulators and targets do exist and the implications are significant in order to understand the cellular behavior (Wagner and Wright 2007). The systematic computational identification of putative regulatory structures would therefore enable a more detailed analysis. Within an optimization framework however, such alternative structures can be identified and critical nodes whose removal would be lethal can be speculated. Similarly, interchangeable nodes can also be proposed. These sub-optimal alterative structures provide mechanisms by which an organism compensates for changes in environmental conditions. Therefore, while the response may not be optimal, the organism is more flexible and remains viable under a wide range of environmental conditions. One piece of evidence which supports the activation of alterative structures is the continuing viability of different *E. coli* strains despite knockouts of important regulatory proteins. If alternative network architectures do not exist, then the viability of a given strain would be severely compromised. Therefore, we believe that our method of deciphering alternative regulatory networks corresponds biologically to the inherent flexibility exhibited by organisms.

2.3 Analysis of regulatory networks

Deciphering the structure of regulatory networks should be considered as the prelude to further analyses that attempt to elucidate putative roles of the regulators rather than a rigorous and restrictive reconstruction of experimental data. After all, it is widely accepted that multiple, alternative, regulatory networks can reproduce experimental data (Tran et al. 2005). As such, a number of questions emerge, namely: *Can these networks be identified in a systematic and unbiased manner? Are there any persistent interactions that emerge from multiple architectures? Are there specific transcription factors whose activity profiles remain robust across multiple realizations? Can the specific function of the undetermined factors, i.e., factors can act either as activators or repressors, be systematically determined? Do preferential patterns emerge in terms of the nature of these factors i.e., activators or repressors?*

We are proposing a mathematical programming approach to address these questions. The model will be presented in detail in Sect. 2.4. In the next section the issue of nonconvexity of Eq. 4b is first considered.

2.4 Model linearization

As mentioned above the definition of P^{eff} in (4b) introduces a non-convex bilinearity in the formulation due to the product of the continuous variable $P(j,t)$ and the binary variable $r(i,j)$. However, this product is exactly linearized through the introduction of the set of constraints defined in (7). In the case of a repressor the general form reduces to (7a). The second constraint is inactive (M is a big number) whereas the first constraint forces $P^{eff}(i,j,t) = -P(j,t)$. The implication is that because "j" acts a repressor of "i" if the activity of $P(j,t)$ increases, i.e., $P(j,t) > 0$, the effect of $E(i,j,t)$ should be of the opposite sign and therefore result in reduction of $E(i,j,t)$, i.e, $E(i,j,t) < 0$. Similarly, if the activity of $P(j,t) < 0$, because "j" acts as a repressor, then reduction in its activity should enhance the expression of $E(i,j,t)$, i.e., $E(i,j,t) > 0$. When $r(i,j) = 1$("j" acts as an activator of "i") the system reduces to form (7b) which makes the first constraint redundant, whereas the second constraint forces $P^{eff}(i,j,t) = P(j,t)$ and therefore it acts as an activator. N_{TF} is the number of transcription factors, N_g is the number of genes, and N_T is the number of time points.

$$
\begin{aligned}
&\text{general form}\\
&-r(i,j)\cdot M - P(j,t) \le P^{eff}(i,j,t) \le r(i,j)\cdot M - P(j,t)\\
&\big[r(i,j) - 1\big]\cdot M + P(j,t) \le P^{eff}(i,j,t) \le \big[1 - r(i,j)\big]\cdot M + P(j,t)\\
&\text{(a) modeling a repressor: } r(i,j) = 0\\
&-P(j,t) \le P^{eff}(i,j,t) \le -P(j,t)\\
&-M + P(j,t) \le P^{eff}(i,j,t) \le \cdot M + P(j,t)\\
&\text{(b) modeling an activator: } r(i,j) = 1\\
&-M - P(j,t) \le P^{eff}(i,j,t) \le M - P(j,t)\\
&P(j,t) \le P^{eff}(i,j,t) \le P(j,t)
\end{aligned}
\tag{7}
$$

2.5 Integer optimization formulation

We propose a mixed-integer formulation able to effectively address the aforementioned questions in a unified framework. The complexity of the regulatory network is controlled through the introduction of a binary variable z(j) which denotes the existence, z(j) = 1, or non-existence of a particular regulator's activity z(j) = 0. It should be emphasized that eliminating the effect of a regulator implies blocking the activity of the TF and not, necessarily, the expression of the corresponding gene. The underlying assumption behind this modeling exercise is to identify what types of alternative structures can be constructed that reproduce optimally the experimental expression data. The complexity of the network is controlled by setting the required number of non-zero elements in this variable set. Furthermore, alternative structures for the same number of transcription factors can be generated by introducing appropriate cuts (8) that exclude previous integer solutions, i.e., combinations of non-zero z(j)'s (Biegler et al. 1997).

$$\begin{aligned}&\sum_{j\in N^k} z(j) - \sum_{j\in B^k} z(j) \le \left|N^k\right| - 1\\ &N^k = \{j|z^k(j) = 1\},\quad B^k = \{j, |z^k(j) = 0\}\end{aligned} \tag{8}$$

In order to identify structurally robust elements of the regulatory architecture we introduce a robustness metric which quantifies the number of times a particular TF appears in each of the alternative structures in conjunction with the robustness of the reconstructed activity profile. The metric is therefore: R(j) = [f(j)/S]*C(j), where R(j) is the robustness of TF "j" when we generate multiple network modules, f(j) describes the frequency of TF j across the multiple solutions S (simply it shows how many times TF j appears in different network architectures), C(j) corresponds to the average Pearson's Correlation coefficient for the multiple inferred activities (P(j, t)) of TF j and M is the total number of alternative structures under consideration.

The optimization framework attempts to deconvolute the gene expression profiles in terms of a reduced "basis set" defined by the activities of the corresponding TFs. The aim is to achieve the best possible decomposition while utilizing prior knowledge about the systems, in terms of known interactions as well as the possibly known directionality of a subset of those interactions (activation/suppression). Furthermore, we are interested in identifying systematically alternative structures in order to unravel the potential underlying structure of the regulatory network by pint pointing robust and, presumably, critical regulators. All of the above questions can indeed be addressed by the solution of the mixed-integer linear optimization problem, miSARN—mixed integer Synthesis and Analysis of Regulatory Networks (9), solved using the GAMS modeling software (Brooke et al. 2004) running CPLEX 9 for the solution of the corresponding MILP. The parameter estimation problem is kept linear by replacing the error term with the positive and negative slacks e^+ and e^- which are subsequently minimized.

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

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

subject to

$$E(i,t) - \sum_j \pi(i,j)P^{eff}(i,j,t) = e^+(i,t) - e^-(i,t) \quad \forall i,t$$

$$
\begin{aligned}
&\sum_j z(j) = m \le N_{TF}\\
&\sum_j D(i,j)\cdot z(j) \ge 1 \quad \forall i\\
&-r(i,j)M - P(j,t) \le P^{eff}(i,j,t) \le r(i,j)M - P(j,t) \quad \forall i,j,t\\
&[r(i,j){-}1]M + P(j,t) \le P^{eff}(i,j,t) \le [1 - r(i,j)]M + P(j,t) \quad \forall i,j,t\\
&z(j)P\min \le P(j,t) \le z(j)P\max \quad \forall j,t\\
&\sum_{j\in N^k} z(j) - \sum_{j\in B^k} z(j) \le \left|N^k\right| - 1 \qquad (9)\\
&N^k = \{j|z^k(j) = 1\}, \quad B^k = \{j, |z^k(j) = 0\}\\
&D(i,j) = \begin{cases} 1 & \pi(i,j) \ne 0\\ 0 & \pi(i,j) = 0\end{cases} \quad \forall i,j\\
&P(j,t), P^{eff}(i,j,t) \in \Re\\
&e^+(i,t), e^-(i,t) \in \Re^+ \quad \forall i,j,t\\
&z(j), r(i,j) \in \{0,1\} \quad \forall i,j\\
&i = 1,\ldots,N_g; \quad j = 1\ldots,N_{TF}; \quad t = 1\ldots N_T
\end{aligned}
$$

3 Results

Temporal expression profiles of *E. coli* during transition from glucose to acetate as the sole carbon source were detected using DNA microarrays. The importance of such experiment lies on the premise that glucose and acetate are utilized by distinct metabolic pathways and thereby understanding such profiles in different carbon sources gives us a more thorough insight about the dynamic behavior of *E. coli* (Oh et al. 2002). The temporal *E. coli* expression data as well as the connectivity matrix for this system are publicly available at http://www.seas.ucla.edu/~liaoj/. The data included the log transformed expression levels (relative to initial time point) of 100 genes recorded at 10 time points. Such expression data have been part of studies (Kao et al. 2004; Boulesteix and Strimmer 2005; Pournara and Wernisch 2007). The corresponding connectivity matrix given the available information of RegulonDB (Salgado et al. 2001) database. Based on RegulonDB information we fix the binary variables r(i, j) to be either 0 or 1 if j is known to repress or activate gene i, respectively. All others are treated as variables whose type of regulation will be determined based on the solution of the optimization problem (predicting the regulatory role of three transcription factors that are known to regulate six target genes). Therefore, the final MILP formulation contains 37 binary variables including 30 which determine whether a given transcription factor is used, and 7 which characterize the nature of the interaction (activator/repressor) for the undetermined pairs, as explained in greater detail in Sect. 4.

3.1 Systematic generation of alternative regulatory structures

The complete regulatory structure is composed of 30 transcription factors. Given the hard constraint that each gene must be regulated by at least one TF, the miSARN formulation becomes infeasible if less than 18 factors are used since this many factors are needed to guarantee that all genes are properly regulated, that is there is no combination of less than

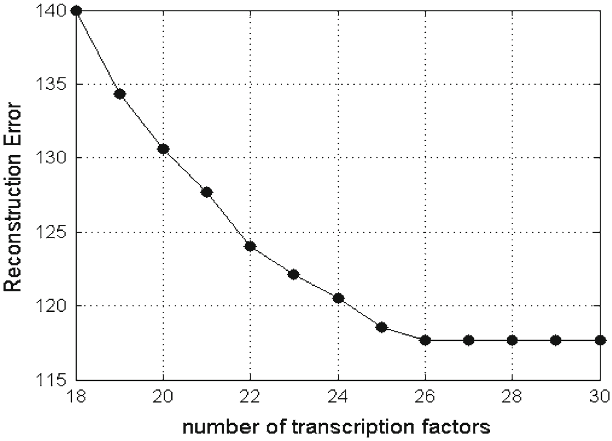

Fig. 2 Reconstruction error as a function of the number of active transcription factors

18 TFs that would make sure each gene is regulated by at least one factor. Varying the control parameter "m" in the range of 18–30 TF generates an equivalent non inferior set as shown in Fig. 2. Interestingly we observe that there are five different network architectures (m = 26…30) that generate architectures resulting in the same reconstruction error, despite the fact that each utilizes a different number of TFs.

Given the availability of these alternative structures, we proceed to evaluate the robustness of each factor across multiple solutions. The results are summarized in Table 1. It is clear that a critical subset emerges that not only persist as a selection of active TF, but also the corresponding reconstructed profiles are very robust across multiple solutions. The reconstructed profiles for all factors across all the 13 solutions (m = 18…30) are depicted in Fig. 3. For each of the cases that result in the lowest reconstruction error (m = 26, 27, 28 and 29) we evaluate the number of alternative structures for each value of m that generate networks with the same reconstruction error. This is achieved by activating the integer cuts that eliminate the previous integer optimal solution. The miSARN formulation identifies 8 alternative structures for m = 29, 24 structures for m = 28, 32 structures for m = 27 and 16 structures for m = 26. In total 80 alternative structures with different number of factors for each family and different connections are identified that result in the same approximation error. The implications of the robust selection as well as the alternative architectures are discussed in the following section. Typical reconstructed expression profiles are provided in Fig. 4.

4 Discussion

Analysis of the results of Table 1 provides us with a list of putative critical regulators, which are characterized by robust activity profiles. Concentrating on the sub-set of regulators that correspond to R(j) = 1, we identify the following factors: *Ada, CysB, FadR, GatR, LeuO, Lrp, PurR, TrpR*, and *TyrR*.ll to play a critical role in the metabolism of *E. coli* when carbon source transition occurs from glucose to acetate. In particular, *Ada* regulates aidB, which belongs to the adaptive response genes and encodes for a protein that is homologous to mammalian acyl coenzyme A dehydrogenases. This activation is crucial during either anaerobiosis state or acetate metabolism (Landini et al. 1994). Moreover, all the other regulators influence

Table 1 Robustness index for all transcription factors

TF name	Relative connectivity	f(j)	C(j)	R(j)
Ada	1	13	1.0	1.0
CysB	4	13	1.0	1.0
FadR	3	13	1.0	1.0
GatR	4	13	1.0	1.0
LeuO	3	13	1.0	1.0
Lrp	6	13	1.0	1.0
PurR	3	13	1.0	1.0
TrpR	3	13	1.0	1.0
TyrR	6	13	1.0	1.0
ArcA	18	13	0.9	0.9
PhoB	5	13	0.9	0.9
FIS	7	11	1.0	0.9
NarL	9	13	0.9	0.9
CRP	21	13	0.9	0.9
RpoE	8	13	0.9	0.9
RpoS	5	13	0.7	0.7
FruR	7	13	0.7	0.7
OmpR	3	13	0.6	0.6
IHF	12	13	0.6	0.6
IclR	4	12	0.9	0.6
GlpR	1	8	1.0	0.5
LexA	1	5	1.0	0.4
PspF	1	5	0.6	0.4
FNR	16	10	0.4	0.2
CsgD	3	2	1.0	0.2
Rob	3	1	1.0	0.2
SdiA	1	5	0.2	0.1
RpoN	1	8	0.2	0.1
GalR	3	7	0.0	0.0
RcsAB	1	4	0.0	0.0

the expression of crucial metabolic genes necessary during the specific growth arrest. *CysB* regulates genes essential to sulfur utilization and nitrogen metabolism whereas *GatR* regulates genes important to galactol utilization and transport. In addition to this, *PurR* is a key repressor protein for purine nucleotide synthesis and it is likely to coregulate other genes for de novo purine nucleotide synthesis (Rolfes and Zalkin 1988). Meanwhile, *FadR* is characterized as a global regulator in fatty acid biosynthesis and degradation (DiRusso et al. 1992) and the leucine responsive regulatory protein (*Lrp*) is another global regulator of metabolism in *E. coli* (Calvo and Matthews 1994). Furthermore, *GatR* regulates genes essential to galacticol utilization and transport and in Chen and Wu (2005) it is emphasized that *LeuO* is characterized by a gene silencing activity. Such activity is integral to the regulation of prokaryotic and eukaryotic gene expression and given that we are allowed to “target” such transcription factors we are closer to unraveling the underlying complexities of gene regulation. Regulators such as *TrpR* and *TyrR* are characterized as major transcription regulators for a group of genes that are essential for aromatic amino acid biosynthesis and transport in *E. coli* (Lawley and Pittard 1994; Lawley et al. 1995). Equally important is the analysis of the robustness characteristics of the functional charctaerization of the regulators whose activity (activator/repressor) is not uniquely determined. Out of the 30 TFs three have been experimentally assigned a dual function (act as either activator or repressor). These TFs along

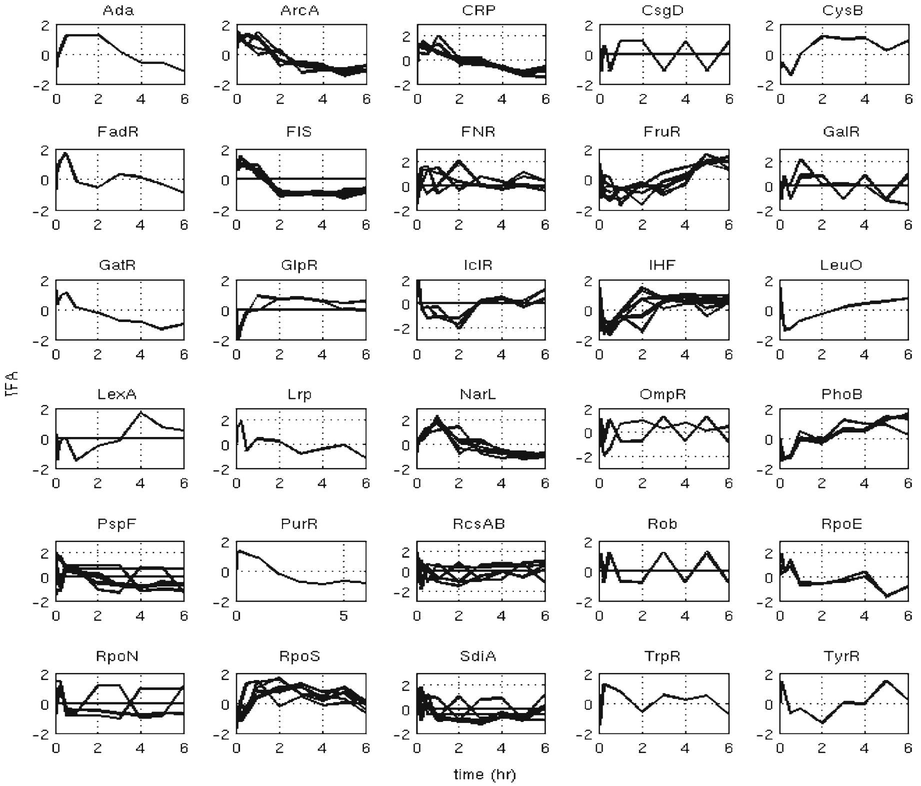

Fig. 3 Reconstruction of TFA profiles

with their target genes are: (1) *CRP*: galE, galK, galT, prop; (2) *LRP*: kbl; (3) *PhoB*: ugpB, ugpE. The inferred role of the 3 dual TFs across the 13 multiple solutions is robust for all the dual 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.

The incorporation of the cuts excluding previous solutions for a given value of m and the generation of the alternative structures generates equally interesting results. The multiple architectures for m = 29 effectively define networks in which one TF is eliminated from the network (Fig. 5). There are four distinct modules that give rise to these solutions and all cases effectively amount to the elimination of the activity of a factor provided that its contribution can be represented by another factor. The interchangeable pairs are: (*PspF*, *RpoN*), (*SdiA*, *RcsAB*), (*CsgD*, *OmpR*), and (*Rob*, *GalR*). These alterative structures may prove to be important in explaining the viability of different strains of *E. coli* as well as its ability to tolerate a variety of environmental conditions while still retaining its viability. Therefore, these alternative structures are important aspects of the network and allow us to separate the vital connections from those that impart flexibility. These findings, albeit computational, can be characterized as both challenging and promising on the premise that there is on-going research about identifying clinically intervention points whose effective combinatorial inhibition would improve the process of therapeutic drugs. There are several studies

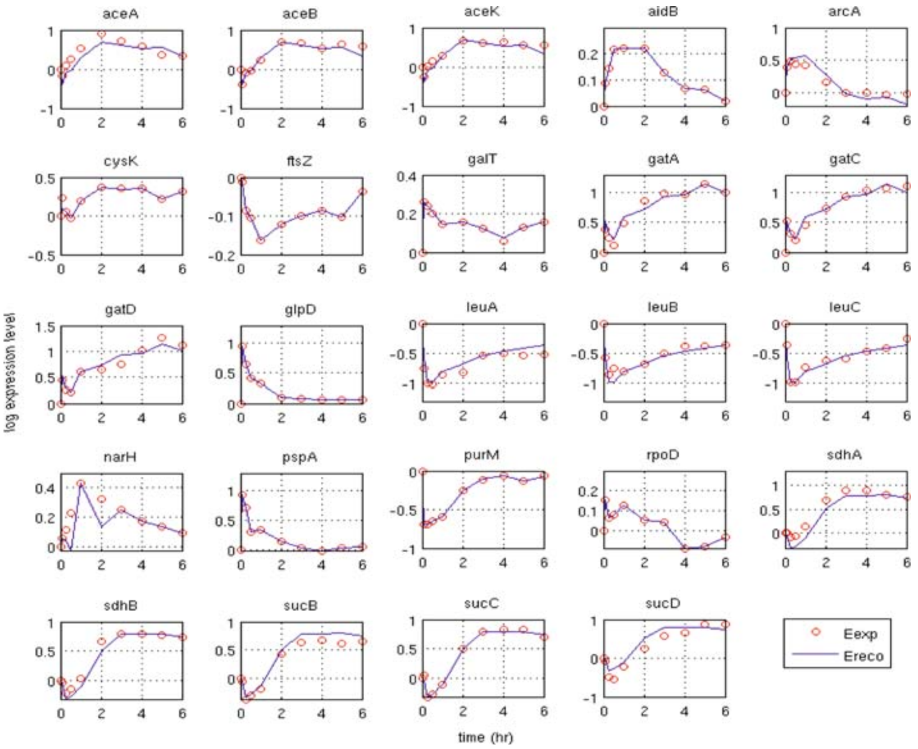

Fig. 4 Reconstruction of gene expression profiles

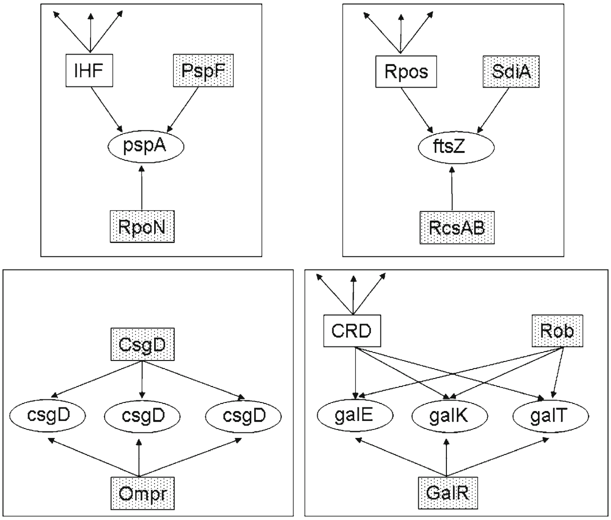

Fig. 5 Alternative equivalent regulatory structures. Square: TF; oval: genes; dark squares: interchangeable TFs. CsgD denotes the activity of the corresponding TF, csgD denotes the gene

(Covert et al. 2004; Kato et al. 2004) that seek to unravel the underlying principles that govern gene regulation by either combining sequence data with binding data such as Chip-chip data and expression data or by knocking out (deleting) transcription factors and binding sites with the goal of revealing more about functional regulatory interactions and pathways. The Phage-Shock-Protein System shown in the upper left is regulated by PspF and RpoN promoters in *Y. enterocolitica*, a bacteria very similar to *E. coli*, and 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. For the regulators of ftsZ, it was found that mutants in rcsB which is part of the rcsAB complex had very little difference from that of the wild type strain under normal growth conditions (Gervais et al. 1992). However, despite the similarities under normal growth conditions, it was hypothesized that under different environmental conditions, the presence of a functional rcsB protein may alter the overall response. Additionally, it was found that while the over-expression of sdiA would increase the expression of ftsZ, the deletion of sdiA like the mutations of rcsB did not alter the ability of the cell to divide, and appeared relatively normal under standard conditions (Wang et al. 1991). This is similar to the properties of rcsAB. However, what has not been examined is whether a mutant in both of the sidA and rcsB genes would lead to a significant change in the overall behavior of the organism understand conditions. Similarly, it was found that GalR transcription factor was not necessary under rich growth conditions (Chapuy-Regaud et al. 2003). Currently, there are no studies concerning the null Rob mutants under normal growth conditions, though it was found that under minimal medium conditions such as glucose starved medium that the lack of the Rob transcription factor alters the behavior of *E. coli*, though sub-lethally (Kakeda et al. 1995). One of the limitations in our formulation is that the structure in the lower left of Fig. 5 can be obtained in which CsgD and OmpR are interchangeable. When this structure is given as a directed acyclic graph, the symmetry breaks down for OmpR is found to be a regulator of the transcription factor CsgD, and CsgD autoregulates. Therefore, even though the structures appear to be equivalent, they are not. In spite of the shortcomings in the representation, by generating multiple solutions, and examining the graphs that arise, such inconsistencies can be post-processed into a network representation. The experimental evidence that the mutant strains are indistinguishable from the wild type strains under normal growth conditions validates the computational results which indicated that the error derived from the alternative structures is equivalent. Our results suggest that the removal of both transcription factors would cause a large difference in the error. Therefore it suggests that if there were double knockouts of both of these transcription factors, there would be significant changes in the overall response of the organism. It must be stressed, that while these links show no effect under normal growth conditions, many of these links are significant under different environmental conditions and therefore function to provide flexibility to the organism in the face of changing environmental factors. In addition to the ability to change the overall gene expression profiles, our framework also allows us to easily fix the overall activity of a regulator or remove a regulator depending on conditions which alter the ability of a transcription factor to be activated. Therefore, we assert that in addition to its ability to quantify the strength of the interactions, our framework also has the ability to determine the existence of necessary regulatory structures.

5 Conclusions

Our results demonstrate how an optimization-based model (MILP formulation) can provide us with meaningful biological insights on gene regulation. Our model integrates high-throughput data, network connectivity information as well as known directionality in regulation for specific regulatory pairs, with the aim to reveal underlying principles of the network architecture. Our model can provide both optimal reconstructions and multiple alternative network architectures. We further introduce a metric to distinguish a subset of critical transcription factors, which coupled with a system of integer cuts, provides us with the combinatorial solution of deleting transcription factors. Our model incorporates prior biological knowledge in terms of the effective role of a transcription factor as a regulator or repressor whilst it can decipher the directionality of those TFs that their regulatory role is unknown.

Acknowledgements 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

Alter, O., Golub, G.H.: 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), 16577–16582 (2004)

Biegler, L.T., Grossmann, I.E. et al.: Systematic Methods of Chemical Process Design. Prentice Hall (1997)

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

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

Bussemaker, H.J., Li, H.: et al.: Regulatory element detection using correlation with expression. Nat. Genet. **27**(2), 167–171 (2001)

Calvo, J.M., Matthews, R.G.: The leucine-responsive regulatory protein, a global regulator of metabolism in Escherichia coli. Microbiol. Rev. **58**(3), 466–490 (1994)

Chapuy-Regaud, S., Ogunniyi, A.D. et al.: RegR, a global LacI/GalR family regulator, modulates virulence and competence in *Streptococcus pneumoniae*. Infect. Immun. **71**(5), 2615–2625 (2003)

Chen, C.C., Wu, H.Y.: LeuO protein delimits the transcriptionally active and repressive domains on the bacterial chromosome. J. Biol. Chem. **280**(15), 15111–15121 (2005)

Chen, K.C., Wang, T.Y. et al.: A stochastic differential equation model for quantifying transcriptional regulatory network in *Saccharomyces cerevisiae*. Bioinformatics **21**(12), 2883–2890 (2005)

Covert, M.W., Knight, E.M. et al.: Integrating high-throughput and computational data elucidates bacterial networks. Nature **429**(6987), 92–96 (2004)

DiRusso, C.C., Heimert, T.L. et al.: Characterization of FadR, a global transcriptional regulator of fatty acid metabolism in *Escherichia coli*. Interaction with the fadB promoter is prevented by long chain fatty acyl coenzyme A. J. Biol. Chem. **267**(12), 8685–8691 (1992)

Drazinic, C.M., Smerage, J.B. et al.: Activation mechanism of the multifunctional transcription factor repressor-activator protein 1 (Rap1p). Mol. Cell. Biol. **16**(6), 3187–3196 (1996)

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

Gervais, F.G., Phoenix, P. et al.: The rcsB gene, a positive regulator of colanic acid biosynthesis in *Escherichia coli*, is also an activator of ftsZ expression. J. Bacteriol. **174**(12), 3964–3971 (1992)

Harbison, C.T., Gordon, D.B. et al.: Transcriptional regulatory code of a eukaryotic genome. Nature **431** (7004), 99–104 (2004)

Iyer, V.R., Horak, C.E. et al.: Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature **409**(6819), 533–538 (2001)

Kakeda, M., Ueguchi, C. et al.: An *Escherichia coli* curved DNA-binding protein whose expression is affected by the stationary phase-specific sigma factor sigma S. Mol. Gen. Genet. **248**(5), 629–634 (1995)

Kao, K.C., Yang, Y.L. et al.: Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis. Proc. Natl. Acad. Sci. U.S.A. **101**(2), 641–646 (2004)

Kao, K.C., Tran, L.M. et al.: A global regulatory role of gluconeogenic genes in *Escherichia coli* revealed by transcriptome network analysis. J. Biol. Chem. **280**(43), 36079–36087 (2005)

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

Landini, P., Hajec, L.I. et al.: Structure and transcriptional regulation of the *Escherichia coli* adaptive response gene aidB. J. Bacteriol. **176**(21), 6583–6589 (1994)

Lawley, B., Pittard, A.J.: Regulation of aroL expression by TyrR protein and Trp repressor in *Escherichia coli* K-12. J. Bacteriol. **176**(22), 6921–6930 (1994)

Lawley, B., Fujita, N. et al.: The TyrR protein of *Escherichia coli* is a class I transcription activator. J. Bacteriol. **177**(1), 238–241 (1995)

Lee, T.I., Rinaldi, N.J. et al.: Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science **298**(5594), 799–804 (2002)

Maxson, M.E., Darwin, A.J.: Multiple promoters control expression of the *Yersinia enterocolitica* phage-shock-protein A (pspA) operon. Microbiology **152**(Pt 4), 1001–1010 (2006)

Ng, A., Bursteinas, B. et al.: pSTIING: a 'systems' approach towards integrating signalling pathways, interaction and transcriptional regulatory networks in inflammation and cancer. Nucleic Acids Res. **34**(Database issue), D527–D534 (2006)

Oh, M.K., Rohlin, L. et al.: Global expression profiling of acetate-grown *Escherichia coli*. J. Biol. Chem. **277**(15), 13175–13183 (2002)

Pournara, I., Wernisch, L.: Factor analysis for gene regulatory networks and transcription factor activity profiles. BMC Bioinformatics **8**, 61 (2007)

Rolfes, R.J., Zalkin, H.: *Escherichia coli* gene purR encoding a repressor protein for purine nucleotide synthesis. Cloning, nucleotide sequence, and interaction with the purF operator. J. Biol. Chem. **263**(36), 19653–19661 (1988)

Salgado, H., Santos-Zavaleta, A. et al.: RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. Nucleic Acids Res. **29**(1), 72–74 (2001)

Savageau, M.A.: Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology. Addison-Weslet, Reading (1976)

Stormo, G.D., Fields, D.S.: Specificity, free energy and information content in protein-DNA interactions. Trends Biochem. Sci. **23**(3), 109–113 (1998)

Sun, N., Carroll, R.J. et al.: Bayesian error analysis model for reconstructing transcriptional regulatory networks. Proc. Natl. Acad. Sci. U.S.A. **103**(21), 7988–7993 (2006)

Thomas, R., Mehrotra, S. et al.: A model-based optimization framework for the inference on gene regulatory networks from DNA array data. Bioinformatics **20**(17), 3221–3235 (2004)

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

van Steensel, B., Delrow, J. et al.: Genomewide analysis of *Drosophila* GAGA factor target genes reveals context-dependent DNA binding. Proc. Natl. Acad. Sci. U.S.A. **100**(5), 2580–2585 (2003)

Wagner, A., Wright, J.: Alternative routes and mutational robustness in complex regulatory networks. Biosystems **88**(1–2), 163–172 (2007)

Wang, X.D., de Boer, P.A. et al.: A factor that positively regulates cell division by activating transcription of the major cluster of essential cell division genes of *Escherichia coli*. Embo. J. **10**(11), 3363–3372 (1991)

Wang, W., Cherry, J.M. et al.: A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. U.S.A. **99**(26), 16893–16898 (2002)

Yeung, M.K., Tegner, J. et al.: Reverse engineering gene networks using singular value decomposition and robust regression. Proc. Natl. Acad. Sci. U.S.A. **99**(9), 6163–6168 (2002)