BMC Bioinformatics

Research article

Transcriptional regulatory network refinement and quantification through kinetic modeling, gene expression microarray data and information theory

Abdallah Sayyed-Ahmad[1,2], Kagan Tuncay[1] and Peter J Ortoleva*[1]

Address: [1]Center for Cell and Virus Theory, Department of Chemistry, Indiana University, Bloomington IN 47405, USA and [2]Department of Chemical Engineering and Materials Science, University of Minnesota, 421 Washington Ave SE, Minneapolis, MN 55455, USA

Email: Abdallah Sayyed-Ahmad - asayyeda@cems.umn.edu; Kagan Tuncay - ktuncay@indiana.edu; Peter J Ortoleva* - ortoleva@indiana.edu

* Corresponding author

Published: 23 January 2007

BMC Bioinformatics 2007, **8**:20 doi:10.1186/1471-2105-8-20

Received: 12 July 2006

Accepted: 23 January 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/20

Abstract

Background: Gene expression microarray and other multiplex data hold promise for addressing the challenges of cellular complexity, refined diagnoses and the discovery of well-targeted treatments. A new approach to the construction and quantification of transcriptional regulatory networks (TRNs) is presented that integrates gene expression microarray data and cell modeling through information theory. Given a partial TRN and time series data, a probability density is constructed that is a functional of the time course of transcription factor (TF) thermodynamic activities at the site of gene control, and is a function of mRNA degradation and transcription rate coefficients, and equilibrium constants for TF/gene binding.

Results: Our approach yields more physicochemical information that compliments the results of network structure delineation methods, and thereby can serve as an element of a comprehensive TRN discovery/quantification system. The most probable TF time courses and values of the aforementioned parameters are obtained by maximizing the probability obtained through entropy maximization. Observed time delays between mRNA expression and activity are accounted for implicitly since the time course of the activity of a TF is coupled by probability functional maximization, and is not assumed to be proportional to expression level of the mRNA type that translates into the TF. This allows one to investigate post-translational and TF activation mechanisms of gene regulation. Accuracy and robustness of the method are evaluated. A kinetic formulation is used to facilitate the analysis of phenomena with a strongly dynamical character while a physically-motivated regularization of the TF time course is found to overcome difficulties due to omnipresent noise and data sparsity that plague other methods of gene expression data analysis. An application to *Escherichia coli* is presented.

Conclusion: Multiplex time series data can be used for the construction of the network of cellular processes and the calibration of the associated physicochemical parameters. We have demonstrated these concepts in the context of gene regulation understood through the analysis of gene expression microarray time series data. Casting the approach in a probabilistic framework has allowed us to address the uncertainties in gene expression microarray data. Our approach was found to be robust to error in the gene expression microarray data and mistakes in a proposed TRN.

Background

Gene expression microarray [1-3] and other multiplex data (e.g. NMR, ChIP-on-chip and proteomics) contain a wealth of information, and thereby hold promise for addressing the challenge of cellular complexity and deriving advances in medical sciences [4-7]. Considering the volume of the data and the complexity of the phenomena to be understood, it is evident that the interpretation of such multiplex data must be facilitated by automation. Recently we proposed an approach to the analysis of multiplex bioanalytical data based on the integration of these data with cell modeling through information theory [8]. Here we show how this approach can be extended to the analysis of gene expression microarray time series data.

Kinetic cell models have been used for predicting cell behavior [9-11]. Unfortunately there is a lack of information about many of the rate and equilibrium constants for the reaction and transport processes involved [8,12]. Furthermore, we are presented with the challenge of calibrating and using an incomplete model since key aspects of biochemical networks have yet to be resolved. In contrast, gene expression microarray, protein spectroscopy, NMR, ChIP-on-chip and other multiplex data acquisition techniques yield many simultaneous measurements but they are often only indirectly related to the quantities we seek such as protein and mRNA production and degradation rate coefficients, and TF/gene binding constants, and the stoichiometry of posttranslational processes.

Time series experiments commonly involve monitoring a sample of cells over their cycle or during response to time-varying conditions in the extra-cellular medium such as due to heat shock, transitions to aerobic to anaerobic conditions, from enriched to minimal growth media, or exposure to hormones or drugs. Other dynamical phenomena of interest involve behaviors in response to nuclear transplantation, fertilization or viral infection, as well as the time course of normal development, radiation, transitions to abnormality or drug resistance. Predicting these phenomena, and analyzing of time series data on them can be facilitated using kinetic approaches if the associated dynamic variability is to be explored. In contrast, steady-state approaches can only yield ratios of rate coefficients and not all coefficients independently. Nor can a steady-state approach capture autonomous oscillatory dynamics such as observed during transcription [13,14].

To quantitatively understand the cell, we must account for the omnipresent uncertainty in observed data and in the structure of a cell model. Thus, a probabilistic framework is needed. We suggest that the probability of interest is a function of the rate parameters and initial concentrations and a functional of the time course of the frontier variables for which we do not know the governing equations or experimental measurements. Since we know the time course of gene expression microarray data, in principle, some of the rate parameters, equilibrium constants, initial concentrations as well as the time profile of the frontier variables are more likely to be consistent with it than others. A new approach to the construction and quantification of TRNs is presented here that integrates gene expression microarray time series data and cell modeling through information theory. Given a partial TRN and time series data, a probability density is constructed that is a functional of the time course of TF thermodynamic activities at the site of gene control, and is a function of mRNA degradation and transcription rate coefficients, and equilibrium constants for TF/gene binding.

In attempt to reduce the effect of measurement errors, gene expression microarray data is usually preprocessed via image analysis, statistical approaches and channel normalization before any biochemically viable information is derived [15-17]. A number of methods have been proposed for extracting information and overcoming systematic errors from gene expression microarray data after preprocessing it. Among them are Boolean network models [18], Bayesian network models [19], Bayesian statistics [20,21], cluster analysis [22,23], independent component analysis (ICA) [24,25], principal component analysis (PCA) [26,27] and network component analysis (NCA) [28,29]. These techniques are based on the assumption that the system is at steady-state.

The goal of Boolean model analysis is to infer gene regulatory network structure. However, Boolean network models oversimplify gene expression by using a binary approximation wherein genes are considered either active or inactive. The interaction between genes is then represented by Boolean functions (e.g. AND, OR, etc.), and hence the state of a gene (active/inactive) is calculated using the state of the controlling genes. A regulatory network is then constructed by searching all possible Boolean functions until a network that best fits all the data is obtained. While such approaches miss the subtler variation in the degree of gene activity, their computational efficiency allows them to be applied to large networks.

In Bayesian networks, the expression level of every gene is specified by a random variable. Starting form an *a priori* gene regulatory network, gene expression data and using Bayesian statistics, one can construct the conditional probability of the level of expression for each gene given the expression level of another gene that is assumed to regulate it. This conditional probability is then used to build a Bayesian network by keeping all edges (i.e. assumed regulatory interactions) that have a conditional probability higher than a threshold.

Cluster analysis, Bayesian statistics, ICA and PCA classify genes into groups; genes that have similar expression profiles are assumed to be similarly regulated or share the same biochemical functionality. However, they cannot uniquely predict the TRN as they do not address the role of TFs mediating gene-gene interactions or the effect of external factors (e.g. carbon source or TF activators/deactivators such as hormones). Cluster analysis is based on statistical techniques wherein correlations are sought between the responses of genes. However, the coordination can be extremely complex and circuitous. Thus genes may be part of a multi-branch feedback loop involving several TFs made or activated/deactivated by proteins translated from other genes via a series of kinetic steps that can introduce time delays which can easily mask some interactions or introduce spurious ones. Such effects are even more pronounced in light of noise in the observed expression profiles. Furthermore, for a given gene, there is no established correlation between mRNA expression and the level of protein it translates [30]. These time-delayed, complex relationships are revealed by our method which explicitly accounts for the role and time course of the TFs.

NCA differs from other techniques in that the structure of the TRN is assumed to be known. A number of assumptions are made in NCA to arrive at the final steady-state model. The approach presented here requires at least a part of the TRN. However, we place no restrictions on the structure of this network, use a kinetic model, construct synthetic gene expression microarray time series, apply a physically motivated regularization constraint for the time-dependence of TF activities that enhances robustness, places the entire computation in an information theory context so that the uncertainty can be assessed, and then analyzes TRN structure and the associated physicochemical parameters. The latter include mRNA production and degradation rates and TF/gene binding constants. The use of a kinetic model also allows us to generalize our approach to proteomic and metabolic data either by themselves or with gene expression microarray data.

Our method is significantly different from the approach proposed by Gardener *et al.* [31] whose objective is to construct the gene control network using gene expression microarray data and limiting the number of interactions per gene. However, even when there are just a few interactions per gene, there can be thousands of networks that can explain the same gene expression microarray data with a given accuracy. A variety of other methods has been also proposed for TRN construction and augmentation; these include gene ontology, phylogenetic profiles [32] and promoter sequence analysis [33]. The methodology presented here is meant to compliment other approaches and act as a filter for spurious networks by contrasting predictions with observed expression and predicting TF time courses. The later provides an important framework for TF/gene interactions or a self-consistency check on predictions of the other methods. Furthermore, as other methods suggest that there could be TF/gene interaction, our methodology compliments this by providing the specific nature (up/down) of the regulation.

In this paper, we present our approach and apply it to *E. coli*. Our analysis is based on a simplified approach wherein only TRN structure is obtained. Next the resulting TRN is used with the kinetic/information theory approach presented here to calibrate the physicochemical parameters and refine the network structure.

Methods

Schematically, our approach to the incomplete-model challenge is as follows. The state of a cell is specified by a set of variables $\underline{\Psi}$ for which we know the governing equations and a set $\underline{T}$, which is at the frontier of our understanding (i.e. for which we do not know the governing equations). The challenge is that the dynamics of $\underline{\Psi}$ is given by a cell model, e.g.

$$\frac{d\underline{\Psi}}{dt} = \underline{G}(\underline{\Psi}, \underline{T}, \underline{\Lambda}), \qquad (1)$$

in which the rate G depends not only on many rate and equilibrium constants $\underline{\Lambda}$, but also on the time-dependent frontier variables $\underline{T}(t)$. The descriptive variables, $\underline{\Psi}$, can only be determined as a functional of the unknown time courses $\underline{T}(t)$. Thus the model cannot be simulated.

Gene expression microarray time series data, M, reflects the evolving state of the genome and hence one can compare it with a model predicted synthetic time series, M^{syn}, to derive a measure of the accuracy and completeness of the model. Solving Eq. 1, yields the time-course of $\underline{\Psi}$ over the duration of the experiment as a function of $\underline{\Lambda}$, the initial state $\underline{\Psi}(t = 0)$, and as a functional of $\underline{T}(t)$ (i.e. $\underline{\Psi} = \underline{\Psi}[\underline{T},\underline{\Lambda},\underline{\Psi}_0]$); thus the gene expression microarray data constructed, M^{syn}, when compared with M, yields a measure of the accuracy of $\underline{T}(t)$, $\underline{\Lambda}$ and $\underline{\Psi}_0$. In the present application, $\underline{\Psi}$ represents intracellular mRNA levels, $\underline{\Lambda}$ is the aforementioned set of rate and binding constants, and $\underline{T}(t)$ is the time course of TF activities at the site of gene control.

To quantitatively understand the cell, we must account for the omnipresent uncertainty in observed data and in the structure of a cell model. Thus, a probabilistic framework is needed. We suggest that the probability of interest (denoted ρ) is a function of $\underline{\Lambda}$ and $\underline{\Psi}_0$ and a functional of the time course of the frontier variables $\underline{T}(t)$. Since we know M, in principle, some $\underline{\Lambda}$, $\underline{T}(t)$ and $\underline{\Psi}_0$ are more likely

to be consistent with it than others. We develop a model (e.g. a realization of Eq. 1) and use time series gene expression microarray data with information theory to construct ρ.

In the present approach the physics and chemistry of the mechanisms built into a kinetic cell model puts constraints on the relationship between $\underline{\Psi}$ and $\underline{T}(t)$, $\underline{\Lambda}$ and $\underline{\Psi}_0$; this facilitates the determination of these variables from time series data. In a sense, physics and chemistry of biochemical reaction/transport processes enhance the solvability of the inverse problem for the determination of $\underline{T}(t)$,$\underline{\Lambda}$ and $\underline{\Psi}_0$ given the time series gene expression microarray data. Since we know M, in principle, some $\underline{\Lambda}$,$\underline{T}(t)$ and $\underline{\Psi}_0$ are more likely to be consistent with it than others. We develop a model (e.g. a realization of Eq. 1) and use time series gene expression microarray data with information theory to construct ρ.

A transcription model

Gene expression is a multi-step process that is mainly regulated by proteins that activate or repress transcription. In prokaryotic genes, transcription initiation is controlled by promoters which are DNA sequence elements recognized by RNA polymerase. The activity of RNA polymerase is regulated through the interaction of the DNA-binding proteins known as transcription factors with short, specific DNA sequences. These sequences are normally located close to the promoter of the regulated gene. DNA-binding proteins alter the binding affinity of RNA polymerase, consequently affecting the RNA transcription rate [34]. It is evident that the physiological function of a DNA-binding protein is driven by its binding affinity with a gene promoter or adjacent DNA sequence. In particular, Repressor proteins bind to the promoter site thus competing with RNA polymerase for the same binding site while activator proteins usually bind adjacent to the promoter site and hence enhancing the binding affinity of RNA polymerase.

In our approach, we no longer require a comprehensive cell model to make quantitative predictions even though processes within and among the genome, proteome and metabolome are all strongly coupled. Rather our approach only requires an incomplete model wherein governing equations for the variables $\underline{\Psi}$ of Eq. 1 are to be set forth and the model must contain those variables with which one may construct synthetic data of the type available. Thus the following model is based on equations for mRNA levels so that synthetic gene expression microarray data can be constructed. But to do so one must know the time course of TFs as they up/down regulate genes. Specifically, to implement our approach the TF intra-nuclear thermodynamic activities are identified as the frontier variables $\underline{T}(t)$ of (Eq. 1). Closure is obtained by using time series data to generate functional differential equations for the $\underline{T}(t)$. Thus one need not make oversimplified assumptions on $\underline{T}(t)$ to compute $\underline{\Psi}$ (i.e. one may calibrate and run an incomplete transcription model even though the dynamical variables (e.g. RNA populations) are strongly coupled to $\underline{T}(t)$).

We first develop a forward model in which given a set of time courses of TF thermodynamic activities, $\underline{T}(t)$, the time course of intra-cellular mRNA populations $\underline{R}(t)$ is predicted. Due to the dense environment within the nucleus, TF thermodynamic activities are preferred over concentrations. With such a model and time series gene expression microarray data, we now show how transcription rate and other parameters, the most probable $\underline{T}(t)$ can be obtained, and the gene control network can be quantitatively characterized. Finally, note that the following model is rather simple. While more complete models could be used [11], our purpose here is to demonstrate the approach and not to address the supercomputer challenge of a very detailed approach.

Given a TRN with N_{TF} TFs and N_g genes, the i-th gene (g_i), $i = 1,\cup N_g$, is assumed to have $N_s^{(i)}$ uncompetitive TF binding sites labeled $j = 1,\cup N_s^{(i)}$. Assume the binding at any site is independent of the state of others and that only one type of TFs can bind to each site. While the latter two assumptions are not inherent restrictions of our approach, they are made here for simplicity. Let n_{ij} label the TF that can bind to site j on gene i. Let b_{ij} be minus one or plus one indicate the nature of regulation (down or up) of TF type n_{ij} on g_i. At TF/gene equilibrium, the probability H_i that g_i is available for RNA polymerase (RP) complexing is taken to be given by an equilibrium Langmuir uncompetitive adsorption isotherm [35]

$$H_i = \prod_{j=1}^{N_s^{(i)}} \left(Q_{ij} T_{n_{ij}} \right)^{\left(\frac{1+b_{ij}}{2} \right)} \Big/ \left(1 + Q_{ij} T_{n_{ij}} \right), \qquad (2)$$

for binding constant Q_{ij} (liter/mole) and intra-nuclear activity $T_{n_{ij}}$ of the n_{ij}-th TF. The rate of RNA transcription initiation is written

$$k_i = k_i^{\max} H_i \left(\underline{\underline{Q}}, \underline{\underline{b}}, \underline{T} \right), \qquad (3)$$

where $k_i^{\max}$ is a saturation rate coefficient that we suggest is diffusion-limited. This assumption is reasonable because it take into account that on one hand, for TF/DNA

and RP/DNA binding, electrostatic interactions tend to lead to higher limiting rate coefficient than the diffusion limited one. On the other hand, the need for having a specific orientation in order for a TF or RP to bind to the DNA tends to lower the limiting rate coefficient than the diffusion limited one [36]. The two aforementioned processes in addition to electrostatic screening due to presence of salt in *vivo* are assumed to balance each other. After RP binds to a gene mRNA elongation commences. If nucleotide concentrations are roughly steady during transcription and the RP advancement velocity u_i (which in principle depends on the sequence of the gene and measured in units of nucleotides/sec), then the transcription polymerization rate A_i is taken to be

$$\frac{1}{A_i} = \frac{1}{k_i[RP]} + \frac{N_{nuc}^i}{u_i}. \qquad (4)$$

A form which captures the rate limiting step of the two serial processes (RP binding and the elongation). With this, the governing equation for the mRNA populations is written as

$$\frac{dR_i}{dt} = A_i - \lambda_i R_i, \qquad (5)$$

where N_{nuc}^i is the total length of the gene g_i, and λ_i is the decay constant. However, for a more detailed model, mRNA degradation could depend on mRNA protein binding factors as well as the level of some hormones or metabolites such as iron [37]. If the rate limiting step for transcription is RNA polymerase binding to the gene [34], then the second term in Eq. 4 may be dropped. Finally, [*RP*] is the activity of free RNA polymerase and is assumed to be constant and henceforth is subsumed in $k_i^{\max}$. The governing equation for mRNA levels evolution becomes

$$\frac{dR_i}{dt} = k_i^{\max} H_i\left(\underline{\underline{Q}}, \underline{\underline{b}}, \underline{T}\right) - \lambda_i R_i. \qquad (6)$$

Finally, our methodology can be generalized by relaxing any of the above assumptions. For example, H_i, can be changed such that competitive binding and TF complexing are accounted for explicitly which in effect will allow for OR logic. Although the extension of our transcription model to include competitive binding is crucial to accurately recover TF activity time courses, this level of description is out of the scope of this study. Further research is needed to obtain specific data on the molecular level about which TF binds to which binding site of a given gene.

Information theory model/data integration

General formulation

Gene expression microarray data is fraught with inaccuracies. Much attention has been placed on minimizing systematic and random errors via quality screening, multi-spot/multi-slide analysis and averaging. Software carrying out these functions yields confidence intervals which are quantitative measures of errors in the experimental data. Information theory was introduced as a method for assessing the uncertainty in the state of a system via an entropy measure [38,39]. In a series of papers [8,40], we have shown how information theory can be used to calibrate model parameters, use an incomplete model and estimate the associated uncertainties based on the inaccuracies in the observed data and the model used. The probability density ρ for the values of the set $\underline{\Lambda}$ of model parameters and the time-dependence of $\underline{T}(t)$ (the set of variables whose governing equations are not in the model, here TF activities within the nucleus) is obtained through entropy maximization.

The development starts with the introduction of the entropy *S*,

$$S = -\int_{\underline{\Lambda},\underline{T}} \mathcal{S}\, \rho \ln \rho, \qquad (7)$$

where $\mathcal{S}$ is an integration over all $\underline{\Lambda}$ and a functional integration over all time courses $\underline{T}(t)$ is indicated. The experimental data and model are introduced via a set of error measures labeled l = 1,2,..., each of whose average $\bar{E}^l$ is assumed known and are given in terms of ρ via

$$\int_{\underline{\Lambda},\underline{T}} \mathcal{S}\, \rho E^l = \bar{E}^l. \qquad (8)$$

For *cDNA* microarray data, $\bar{E}^{cDNA}$ can be estimated from confidence intervals provided by statistical data analysis. According to the information theory prescription we construct $\rho(\underline{T}(t),\underline{\Lambda})$ by maximizing *S* subject to the constraints (Eq. 8), normalization of ρ, qualitative information on the timescale over which $\underline{T}(t)$ can evolve, and other factors reflecting one's expertise. The result is a form for ρ which implies that the $\underline{\Lambda}$, $\underline{T}(t)$ which are most probable yield the lowest error. Also the $\underline{T}(t)$ obtained has no time dependence that is unphysically short. Other constraints could be introduced that allow one to assign higher probability to the range of parameter values $\underline{\Lambda}$ that are near those one expects from experience. Given the inherently subjective nature of probability (e.g. if we know nothing then all $\underline{\Lambda}$, $\underline{T}(t)$ are equally likely), the information theory prescription yields a ρ that is to be consistent with the level of our

knowledge of the system. In our procedure, the resulting ρ is then maximized with respect to $\underline{\Lambda}$ and $\underline{T}(t)$ to determine their most probable values. Cell parameters that must be calibrated to attain a predictive model using our approach are those introduced in the previous section.

Implementation for cDNA microarray data

Our analysis starts with preprocessed data, thus the predictions of our method as other genome wide microarray analysis methods depend on the choice of the preprocessing procedure (although the approach could also be generalized to proceed directly with raw data). Analysis of preprocessed microarray data can be placed in our framework by introducing an associated error E^{cDNA}. Let $M_i\ell$ be the microarray expression level for the i-th of N_g genes in the ℓ-th of N_{micro} experiments (e.g. time slice). Then

$$E^{cDNA} = \sum_{\ell=1}^{N_{micro}} \sum_{i=1}^{N_g} \left(\ln m_{i\ell}^{syn} - \ln m_{i\ell}^{obs} \right)^2, \qquad (9)$$

where $m_i\ell = M_i\ell / M_{iA}$ with A being the initial time or the standard condition. Here, $m_{i\ell}^{syn}$ is the synthetic microarray data constructed from mRNA levels predicted from a cell model (e.g. here that of the previous section), while *obs* indicates an observed value. Thus E^{cDNA} is a function of the set of model parameters $\underline{\Lambda}$ as contained in a cell model (e.g. $\underline{\underline{Q}}, \underline{k}, \underline{\lambda}$ and $\underline{\underline{b}}$). Similarly E^{cDNA} is a functional of the time course $\underline{T}(t)$ of intra-nuclear TF activities. Following the above information theory formulation we introduce the probability $\rho = \rho\,(\underline{T}(t), \underline{\Lambda})$, a functional of the time course $\underline{T}(t)$ and a function of $\underline{\Lambda}$. We construct ρ by maximizing the entropy subject to estimates of the average error measures (here E^{cDNA}) and other information.

The number of time points is restricted due to cost. This fact and the high level of uncertainty in microarray data suggest that the probability functional method cannot yield a meaningful $\underline{T}(t)$ unless more information is known. In our formulation this is introduced via a homogenization constraint that eliminates unphysically short timescale variations in $\underline{T}(t)$ that the sparseness of the time series data would otherwise allow. In particular, we impose the constraint

$$\underset{\underline{\Lambda},\underline{T}}{\mathcal{S}} \rho \int_0^{t_f} \left(\frac{d\underline{T}_n}{dt} \right)^2 dt = t_f \big/ \bar{Q}^2 t_c^2, \qquad (10)$$

for a time series run over the interval from 0 to t_f; t_c is the shortest characteristic time over which we expect that $\underline{T}(t)$ can change appreciably and we assume $\bar{Q}$ (the average TF/gene binding constant) is the inverse of the typical variation of $\underline{T}$.

One can also use a steady-state approximation for information available about post-translational reactions to further constrain S. If a subnetwork of genes with stoichiometric matrix of processes $\underline{d}_n$ is responsible for the production of T_n, then the associated error measure for these processes is

$$E^{TF} = \sum_{n=1}^{N_{TF}} \sum_{k=1}^{N_{times}} \left(T_n(t_k) - \alpha_n \prod_{j=1}^{z_n} m_{d_{nj}}^{obs}(t_k) \right)^2, \qquad (11)$$

where N_{times} is the number of discretized times at which the TF activity is computed, z_n is the number of genes involved in the production of T_n, and α_n is an equilibrium constant. $m_{d_{nj}}^{obs}$ is the observed microarray for the j-th gene responsible for the creation of the n-th TF.

Maximization of the entropy with respect to ρ gives

$$\ln \rho = \ln \Xi - \beta_1 E^{cDNA} - \beta_2 E^{TF} - \omega \sum_{n=1}^{N_{TF}} \int_0^{t_f} dt \left(\frac{dT_n}{dt} \right)^2, \qquad (12)$$

where Ξ is a normalization constant and β_1, β_2, ω are Lagrange multipliers. The multipliers are determined by insuring that the constraints are satisfied. With this the most probable values of $\underline{\Lambda}$, $\underline{T}$, given the microarray data, are obtained by solving $\partial\rho/\partial\underline{\Lambda} = 0$ coupled to $\delta\rho/\delta\underline{T} = 0$, a functional differential equation for $\underline{T}(t)$ that we solve numerically (see appendix A). A discussion of a symmetry rule that applies to the invertibility of microarray data is given in appendix B and implies the need for a minimal amount of regulatory information in order to obtain a unique network. Figure 1 illustrates our approach where microarray and *a priori* TRN is used to infer TF activity time courses and TF/gene binding constants.

Availability and requirements

Project name: **KA**ryote **G**ene **AN**alyzer (KAGAN)

Project home page: https://systemsbiology.indiana.edu/trnd

Operating system(s): Windows (2000 and later versions) or Linux

Programming Languages: F77, php

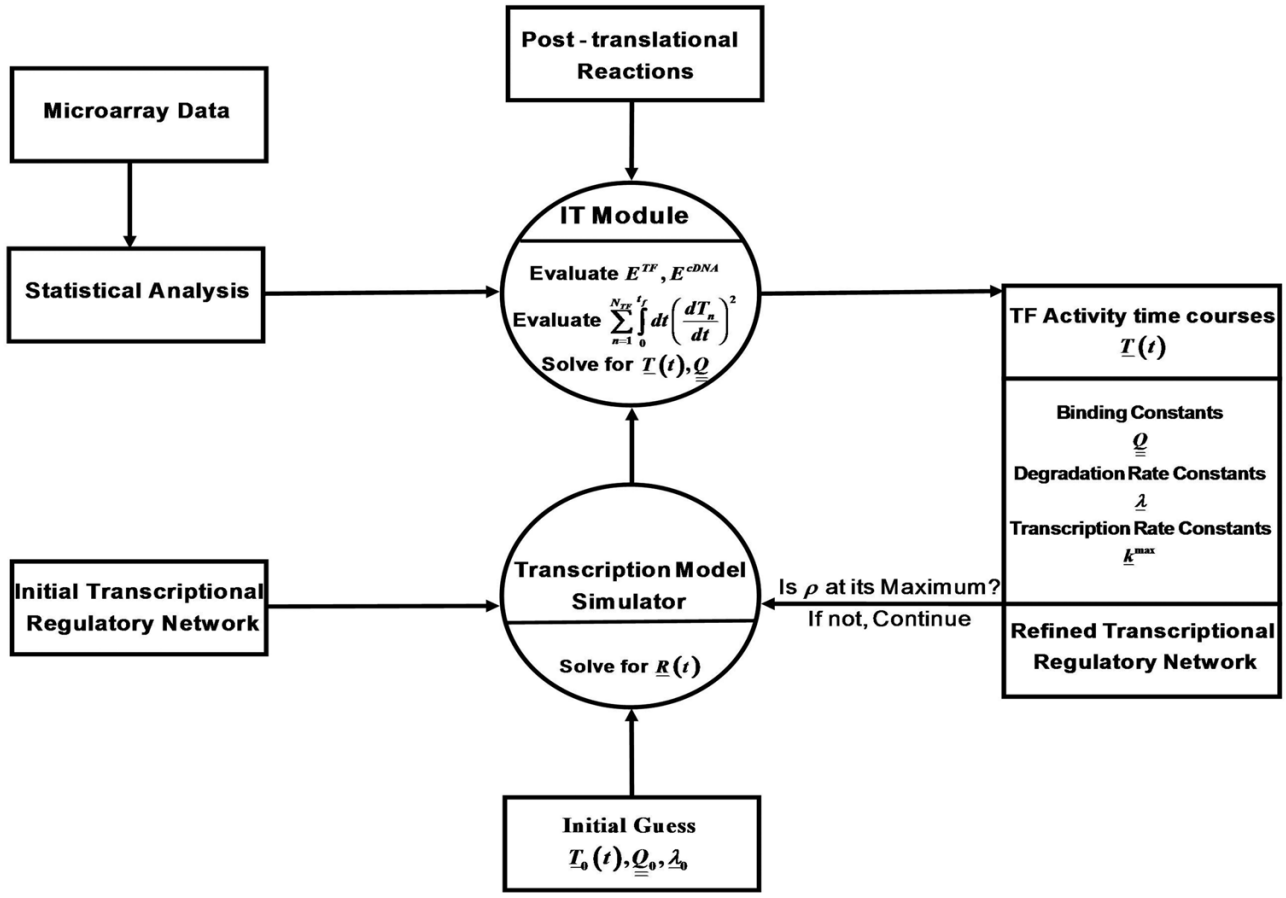

Figure 1
A flowchart of our transcription model/microarray data integration.

Licence: a web interface that allows users to simulate their own data is available from the above site (free registration required).

Restrictions to non-academics: None

Results and discussion

Synthetic example

To test our implementation of the approach described above, and to find its practical limitations, we used a model network that consists of 20 genes and 10 TFs. None of the 20 genes is assumed to code for any of the 10 TFs. The TRN is shown in Table 1 where ± 1 implies up/down regulation. We took all binding constants and initial mRNA concentrations to be unity and 10^{-9} *M*, respectively.

We generated TF time courses according to the following

$$T_n(t) = 1 - 0.5\sin(\upsilon_n t + \varphi_n), \quad (13)$$

where υ_n, φ_n are randomly chosen period and phases. Then we created the synthetic time series microarray data using our transcription model, selecting 10 "data points" that are 500 seconds apart. In the following, we demonstrate the robustness of our approach in reconstructing $\underline{T}(t)$ despite mistakes in the regulatory network and noise in the microarray data, conditions commonly encountered in practice.

Uncertain regulatory network information

Promoter sequence analysis can be used to determine the structure of the TRN based on likely binding sites. However, this approach is likely to suggest a large number of false positive interactions in the TRN. It is of interest to test whether our approach can filter the redundant nonzero entries in the control network. In our approach, if a TF is assumed to upregulate a gene, large binding constants, i.e. $QT \gg 1$, imply that this interaction is unlikely (redundant) as $QT/(1 + QT) \approx 1$. A similar argument hold for wrongly assumed down regulation as indicated by

Table 1: The TRN used for the synthetic example.

	T1	T2	T3	T4	T5	T6	T7	T6	T9	T10
G1	-1	0	0	0	0	0	0	0	0	-1
G2	0	-1	0	0	0	0	0	0	0	1
G3	0	0	-1	0	0	0	1	0	0	0
G4	0	0	0	-1	0	0	0	0	0	0
G5	0	0	0	0	-1	0	0	0	0	0
G6	0	0	0	0	0	-1	0	0	0	0
G7	0	0	0	-1	0	0	-1	0	0	0
G8	0	0	0	0	0	0	0	-1	0	0
G9	0	0	0	0	0	0	0	0	-1	0
G10	0	0	0	0	0	-1	0	0	0	-1
G11	-1	0	0	0	0	0	0	0	0	-1
G12	0	-1	0	0	0	1	0	0	0	1
G13	0	0	-1	0	1	0	1	0	0	0
G14	0	0	0	-1	0	0	0	0	0	0
G15	-1	0	0	0	1	0	0	0	0	0
G16	0	0	0	0	0	1	0	0	0	0
G17	0	0	0	-1	0	0	-1	0	0	0
G18	0	0	0	0	0	0	0	1	0	0
G19	0	0	0	0	0	0	0	0	1	0
G20	0	0	0	0	0	1	0	0	0	1

Columns indicate the TFs whereas the rows indicate the genes that they affect (-1 for down regulation, +1 for up regulation, 0 for no interaction)

small binding constants ($QT << 1$, and therefore $1/1 + QT$) $\approx$ 1). Therefore, our methodology filters out incorrect interactions by assigning large/small binding constants for up/down regulation. To check the vulnerability of our approach to such redundant interactions, we added random nonzero factors in the regulatory network, and obtained the "conjectured full regulatory network" as shown in Table 2. As this network is full (i.e. each gene is regulated by all transcription factors), the NCA method fails. In our approach, the match between the predicted and know TF time courses is remarkable even when the "conjectured" full network was used. Figure 2 illustrates the effect of the number of redundant interactions on the mismatch between the predicted and actual TF time courses. The mismatch (relative to the one obtained using the actual network) does not exceed 2 even when the full interaction network is used as a starting point. Thus, our approach effectively filters the gene control network of unnecessary interactions.

Robustness to microarray error levels

Despite advances in the technology, microarray data has considerable levels of error. There have been few systematic analyses of microarray accuracy due to the many technology platforms available. These platforms have many technological variations which affect the accuracy and reproducibility of the measured expression levels of genes. Such variations are due to multiple techniques of making labeled material, various hybridization conditions, different microarray scanners and settings, etc [41]. Other factors that affect reproducibility of microarray experiments are variations in physiological conditions as well as the number of measurements made to improve the signal to noise ratio. Yeu *et al.* [42] estimated the coefficient of variation for non differentially expressed genes to be 12%–14%, and up to 25% for differentially expressed genes across the entire signal range using 10000 element *E. coli* cDNA microarray. Yuen *et al.* [43] reported a median coefficient of variation of 20.2% when they used cDNA triplicate measurements of 47 genes of *E. coli*. Novak et al. carried out an extensive study of *Affymetrix* Gene Chip oligonucleotide arrays using either identical RNA samples or RNA from replicate cultures under similar biological conditions [44]. They reported an overall coefficient of variation of 24.4% for 4377 genes of the IMR90 human cell line when they used 4 measurements on the same mRNA mixture sample. However, the overall coefficient of variation was 19.9% when they used 11 measurement of mRNA obtained from replicate cultures.

In our implementation, we input raw microarray channel data, perform standard channel normalization (based on housekeeping genes determined using ranking of channel intensities and quality filtering for multiple spot and slide data [16]). The resulting confidence intervals constitute prior information about the level of noise in the microarray response. In this test, we investigate the vulnerability of our approach to error in microarray data. We added

Table 2: In order to test our approach for a large number of TF-gene interactions that might be suggested by sequence analysis or uncertain experimental data, we increased the number of interactions systematically by introducing additional random interactions.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
G1	-1	1	-1	1	1	1	1	-1	-1	-1
G2	1	-1	1	-1	-1	1	1	-1	1	1
G3	1	1	-1	-1	-1	-1	1	1	1	1
G4	1	1	-1	-1	1	-1	-1	1	-1	-1
G5	1	1	1	-1	-1	-1	-1	-1	1	-1
G6	-1	-1	1	1	-1	-1	1	1	-1	1
G7	1	-1	1	-1	1	-1	-1	-1	-1	1
G8	1	-1	1	1	-1	-1	-1	-1	1	1
G9	1	1	-1	-1	-1	1	1	-1	-1	1
G10	-1	1	1	1	-1	-1	1	1	-1	-1
G11	-1	-1	-1	1	-1	1	-1	1	-1	1
G12	-1	-1	1	1	1	1	-1	1	-1	1
G13	1	1	1	1	1	-1	1	-1	1	1
G14	1	-1	-1	-1	1	1	-1	1	-1	-1
G15	-1	1	1	-1	1	-1	-1	1	1	1
G16	-1	1	-1	1	1	1	1	-1	-1	-1
G17	1	1	1	-1	-1	1	1	1	-1	1
G18	1	1	1	-1	-1	1	1	1	-1	1
G19	1	-1	1	-1	-1	-1	-1	1	1	1
G20	-1	1	1	1	-1	1	-1	1	1	1

In the original matrix, there are 34 nonzero elements whereas in the augmented full matrix there are 200 nonzero elements. The challenge is to identify the actual operating TRN.

random noise to the synthetic microarray data that was obtained using the assumed TF time courses, and the transcription regulatory network (Table 1) as follows

$$m_i^{obs}(t) = m_i(t) \times (1 + noise \times (2r - 1)), \quad (14)$$

where r is a random number between 0 and 1. *noise* ×100% represents the coefficient of variation or the percentage noise level in the microarray data. Figure 3 shows the mismatch (relative to the TF time course obtained without added noise) as a function of added noise level with and without the regularization constraint (see Eq. 10). The regularization constraint yields a better match at noise levels higher than 30%, providing robustness to the solution of the inverse problem when noisy data is used. More generally, the Lagrange multiplier for the regularization constraint, Eq. 12, should decrease with the width of the confidence interval. It should be noted that these results were obtained with a fixed Lagrange multiplier for the regularization constraint. If error level is believed to be very small, a smaller Lagrange multiplier should be used. We believe that error levels 30% and higher are to be expected in expression data, in particular for low concentration RNAs. This example also illustrates one of the advantages of time series over steady-state data in that the former is less vulnerable to noise due to the use of our regularization constraint.

Healing mistakes in a proposed regulatory network

Often we rely on TF-gene interactions obtained from questionable quality resources. Therefore, it is important that our algorithm is robust to potential sign mistakes in the TRN due to regulatory differences between the cell line of interest and that for which the network was constructed. In this case, we run our code in a discovery mode that searches for network mistakes. We first rank genes

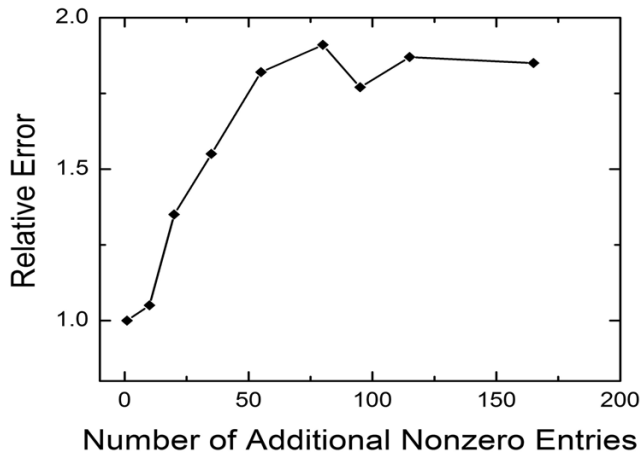

Figure 2
The sum of the square mismatch between the predicted and actual TF time course relative to the one obtained using the actual gene control network shown in Table 1. Although the mismatch increases as the number of interactions in the gene control network increases, it stays within a limit of 2 in this particular example, showing the potential of our approach to discover the operating gene control network hidden in a larger one (see Table 1 and 2 provided in the supplementary material).

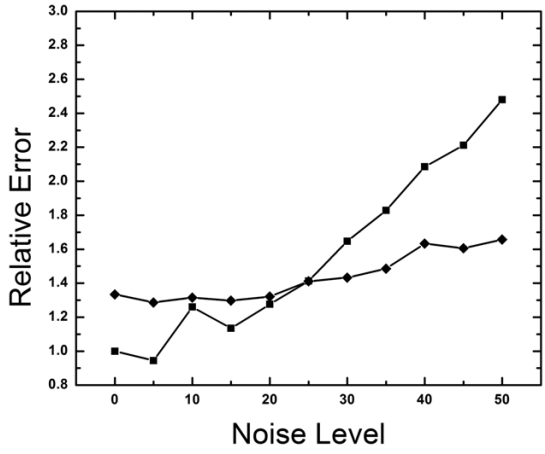

Figure 3
The sum of the square mismatch between the predicted and actual TF time course (relative to that obtained using noise-free microarray data) as a function of the amplitude of random noise added to the microarray data. The diamond and square markers represent the mismatch with and without the regularization constraint. For large noise levels the regularization constraint provides significant improvement in the calibration process.

based on the mismatch between the predicted and observed microarray response, highest ranked having the greatest mismatch. Then, we rank the TFs based on the rank and number of genes that they regulate. As calculation progresses, we periodically check the genes whose mismatch is greater than $E_{average} + a\sigma$ where $E_{average}$ is the average mismatch and σ is the standard deviation of gene mismatch and a is an empirical parameter. Once the genes satisfying this criterion are identified, we change the sign of the regulatory interaction for each of the highest ranked TF (up/down). We also consider additional input that the user provides regarding confidence in each element of the TRN. At a given step in this process, we only change one sign per column of the TRN. After a few iterations, we monitor the mismatch behavior; if the sign change failed to improve the mismatch, we change the sign back. To test this algorithm, we took the TRN of Table 1 and introduced four mistakes by changing the sign of the diagonal elements of genes 1–4. Our algorithm successfully corrected the network after only 10 iterations. Figure 4 shows the predicted and observed microarray data. The triangle markers represent the best fit to the microarray data when the TRN with four mistakes is assumed, whereas the square markers represent the best fit after our algorithm corrected the mistakes. This demonstrates another aspect of our methodology, correcting a user-supplied TRN via microarray data. For large, real systems, we believe that many iterations will be necessary to arrive at an accurate network. The methodology will recommend changes in the TRN based on the microarray data and a user-suggested network. The rankings supplied with the improved network can be used to guide literature searches or carry out sequence analysis that can be used to further refine the network.

Application to E. coli

To test our methodology we used the *E. coli* microarray data obtained for carbon source transition from glucose to acetate media. Details on the experimental conditions and the microarray procedure are provided in Ref. [29]. The data included expression levels (relative to initial time) of 100 genes at 300, 900, 1800, 3600, 7200, 10800, 14400, 18000 and 21600 seconds. The TRN used is based on *RegulonDB*[45] as modified by Kao et al. (2004). We made additional changes based on *EcoCyc* [24]. The final transcriptional regulatory network used is shown in Table 3. Figure 5 shows the time courses of 16 representative TF activities (out of 38). Kao et al. (2004) applied NCA to the same problem. However, the biologically relevant regulatory network that consists of 100 genes and 38 TFs does not satisfy the NCA column rank requirement. Furthermore, the transcription kinetics in our approach differs from the seady-state assumption and binding formulation used in NCA. Despite these differences, 15 out of 16 TF activity time courses (Kao et al. only presented the time courses of 16) are in qualitative agreement. As shown in Figure 5, PhoB increases as a result of response to acetate enrichment of the medium in contrast to the decreasing activity predicted by NCA. Therefore one would expect the phoB activity to increase as well. As a verification of our results, consider the arcA gene which makes ArcA TF that upregulates arcA itself. One would expect a correlation between the expression of arcA and TF activity of ArcA. To have an unbiased test, we took the expression of arcA from the microarray data and calculated the time course of ArcA activity based on the other 17 genes in the network that it regulates. The resultant ArcA time course was indistinguishable from the one obtained by including arcA in the microarray expression data. A comparison of predicted ArcA activity (Figure 5) and expression of arcA Figure 6 shows a similar trend. Figure 6 also shows a comparison between the predicted and observed microarray expression data for nuoJ, nuoA, arcA, livK, ppsA, pykF, pstC and pstS. In a second study, we added up to 30% noise to the *E. coli*. Microarray data, our approach is still found to be robust.

Brown and Callan [33] have predicted many binding sites for the two TF CRP and ArcA. Among the genes included in our model, they predicted that two genes (xthA and livJ) in our data set to be regulated by ArcA. Also they predicted that two genes serA, cyoA and aroP are predicted to be regulated by CRP. To further examine the consistency of our approach with promoter sequence analysis, we performed another simulation after adding these interactions to the TRN obtained form ecoCyc and assuming the nature of these interactions (up vs down) to be unknown. Figure 7 shows improvements in the results obtained for xthA, livJ and serA. Our algorithm predicts that xthA and

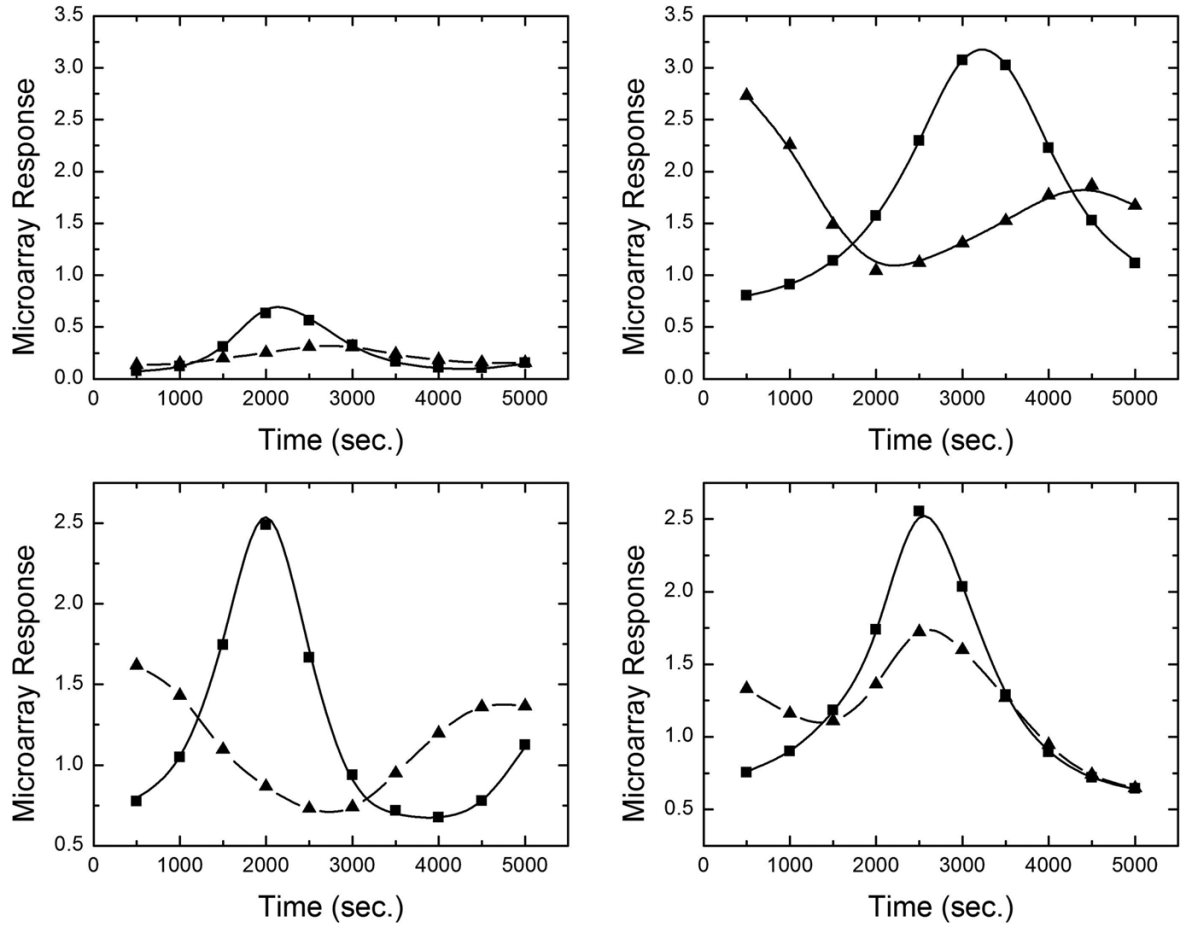

Figure 4
Comparison of observed (solid line) and predicted microarray data for genes 1–4 (see Table 1). Triangles indicate the best microarray fit when there are four mistakes in TRN (the first four entries in the upper left diagonal for genes 1–4); squares indicate the best microarray fit when we allow our program to correct the network. This shows that our algorithm is not only able to calculate the TF time courses, binding constants, etc.; it can also be used as a tool to decide whether a TF is up or down regulating a gene.

livJ are down regulated by ArcA. It also predicts that serA and cyoA are down regulated by CRP. However, no improvement is observed for aroP (Figure 8).

Regularization is important for discriminating between noise/data sparsity-related spurious oscillations and that arising from the nonlinear dynamics of transcription chemical kinetics [13,14]. To demonstrate the effect of regularization, we added 25% noise the observed microarray data. Figure 9, as an illustration, shows that arcA microarray response exhibit oscillatory behavior when no regularization constraint is used. These oscillations are not physical, but rather it is an artifact of using sparse noisy data.

TRN discovery and limitations

A number of issues must be addressed in developing a TRN discovery strategy as follows. Discovering the structure and quantifying the physical chemistry of the gene regulatory network and the underlying mechanism has the challenges that arise in any chemical kinetics problem. For example, the simple process **A+B+C→ABC** can occur through the mechanism **A+B→AB**, **AB+C→ABC** or the other two permutations; to identify the actual mechanism, one must provide intermediate measurements on the dimers **AB**, **BC** and **CA**, rather than simply a measurement of the net rate of **ABC** production. Clearly, in a system with thousands of participating genes and hundreds of TFs, the resolution of the network and the detection of spurious ones, is a grand challenge of combinatorial magnitude. Essentially, all present network discovery methods suffer from this uniqueness difficulty due to the sparsity of available information. In this paper, we demonstrate how our method provides a way to augment an incomplete TRN and to identify inconsistencies in a proposed network based on microarray data.

Table 3: The transcriptional regulatory network used in this study for the 100 responsive genes in E. coli when subjected to carbon source transition from glucose to acetate media.

Gene	Transcription Factor	Gene	Transcription Factor	Gene	Transcription Factor
AceA	-ArcA, -FruR, -IclR, +IHF	ldcC	+RpoS	trpA	-TrpR
AceB	-ArcA, -FruR, -IclR, +IHF	leuA	+LeuO	trpC	-TrpR
AceK	-ArcA, -FruR, -IclR, +IHF	leuB	+LeuO	tyrA	-TyrR
Acs	+Crp, +IclR, +RpoS, +FNR	leuC	+LeuO	tyrR	-TyrR
AdhE	-FruR, -NarL, +RpoS, +FIS	livJ	-Lrp,	ugpB	+Crp, +PhoB
aidB	+RpoS, +ada	livK	+FruR, -Lrp	ugpE	+Crp, +PhoB
arcA	+ArcA, +FNR	lrp	-Lrp, +RpoS, +GadE	uspA	-FadR
aroF	-TyrR	mdh	-ArcA, +Crp, -FlhD	wrbA	+RpoS
aroG	-TyrR	mdoH	+RpoE	xthA	+RpoS
aroM	-TrpR, -TyrR	mutH	+RpoS	yciG	+RpoS
aroP	-TyrR	narH	+NarL, +IHF, +FNR		
csgD	+CsgD, +OmpR	narI	+NarL, +IHF, +FNR		
csgE	+CsgD, +OmpR	narY	+RpoS		
csgF	+CsgD, +OmpR	nrfE	+NarL		
csiE	+Crp, +RpoS, +HNS	nuoA	-ArcA, +NarL, -IHF, -FNR		
cyoA	-ArcA, -FNR	nuoE	-ArcA, +NarL, -IHF, -FNR		
cyoB	-ArcA, -FNR	nuoF	-ArcA, +NarL, -IHF, -FNR		
cysA	+CysB	nuoH	-ArcA, +NarL, -IHF, -FNR		
cysH	+CysB	nuoJ	-ArcA, +NarL, -IHF, -FNR		
cysK	+CysB	osmE	-IHF		
cysM	+CysB	phoR	+PhoB, +TrpR		
dapA	+ RpoE	poxB	+RpoS, +MarA, +SoxS		
Epd	+Crp, +FruR	ppsA	+FruR		
fabA	+FadR	prop	+Crp, +RpoS, +FIS		
FtsZ	+RpoS, +SdiA, +RcsA	pspA	+IHF, +PspF, +RpoN		
gale	+Crp, -GalR, -Rob	pstC	+PhoB		
galK	+Crp, -GalR, -Rob	pstS	+PhoB		
galT	+Crp, -GalR, -Rob	purA	+RpoE		
gatA	-GatR	purK	-PurR		
gatC	-GatR	purM	-PurR		
gatD	-GatR	pykF	-FruR		
gatY	-GatR, +LeuO	rfaF	+RpoE		
glgA	+Crp	rob	+RpoS		
glgP	+Crp	rpoD	+RpoE, -Lexa		
glgS	+Crp	rpoE	+RpoE		
glpD	+Crp, -GlpR	rseC	+RpoE		
gltA	-ArcA, +Crp	sdhA	-ArcA, +Crp, -FNR, -FIS		
Gor	+OxyR	sdhB	-ArcA, +Crp, +NarL, -FNR, -FIS		
guaB	-PurR	serA	+Lrp		
hdeA	+RpoS, +GadX, +GadE	sucB	-ArcA, +Crp, -FNR, -FIS		
hdeB	+RpoS, +GadX, +GadE	sucC	-ArcA, +Crp, -FNR, -FIS		
IclR	-FadR	sucD	-ArcA, +Crp, -FNR, -FIS		
IlvB	+Crp	surA	+RpoE		
ilvH	+Lrp	tnaL	CAP		
Kbl	+Lrp	topA	+FIS		

+ indicates up-regulation
– indicates down-regulation

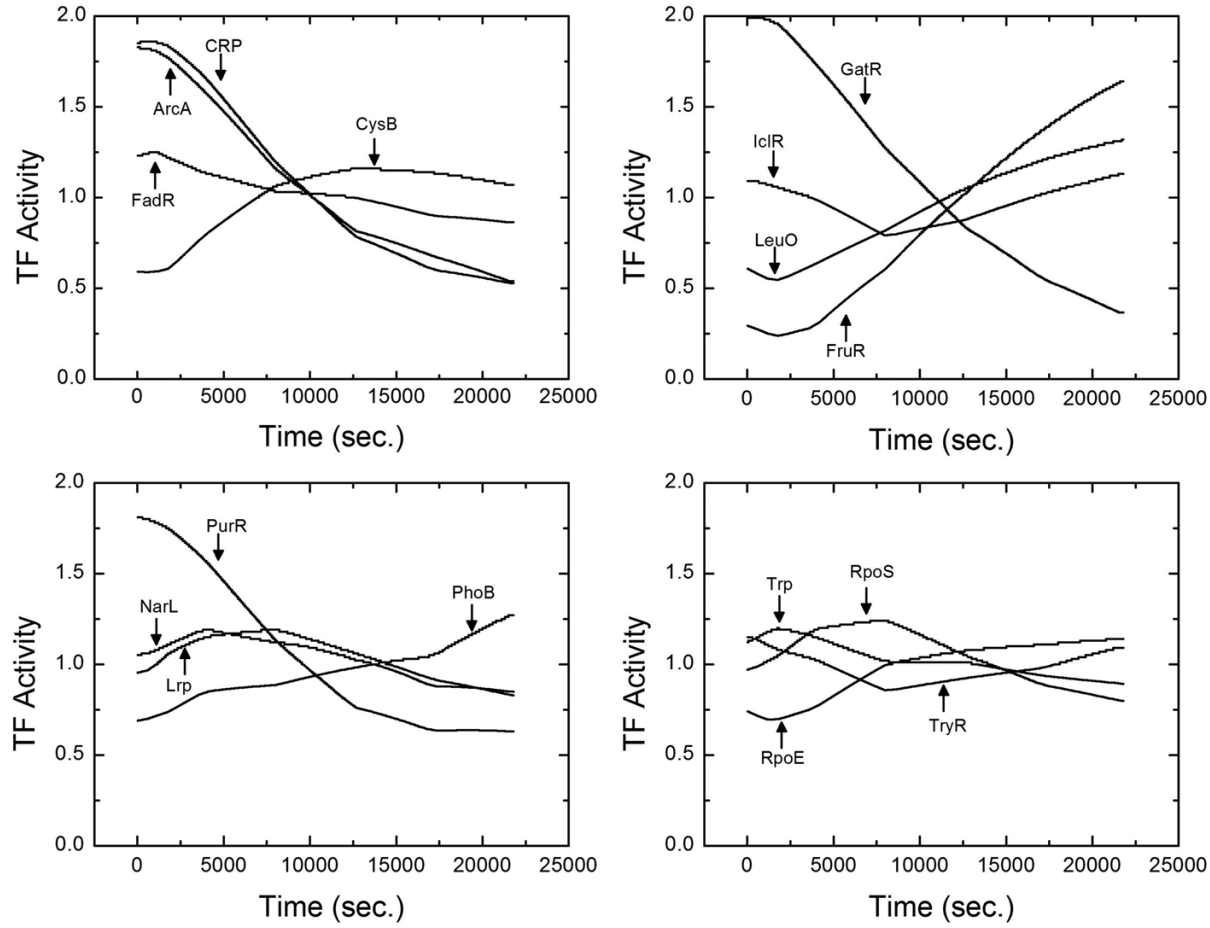

Figure 5
TF activity time courses for 16 of 38 TFs. These results are in qualitative agreement with those obtained by Kao et al. (2003) except for PhoB. pstC and pstS are upregulated by PhoB and their level of expression increases in time (shown in Figure 6), therefore one would expect the activity of PhoB to increase as well.

Processes such as acetylation, methylation and phosphorylation, and the associated enzymes, play important roles in the wider set of pathways [34]. Thus the paradigm **genes→ mRNAs→ proteins → TFs → genes** .... is an oversimplification. While these processes could readily be added to our formulation, it is clear that data in addition to gene expression microarray observations would be required to resolve them. Thus we take the perspective that the simple paradigm cited above can be adopted as a starting point if it is recognized that other processes are somehow mediating the network we quantify. For example, if some genes are repressed in one mammalian cell line by methylation this will be reflected as a small transcriptional rate constant our approach reveals. When predicted levels for a given gene expression are found to be in poor agreement with observations and assuming that the probability that other TFs could be regulating that gene has already been explored, we consider this to imply that the simple paradigm has broken down and the other processes must be acting in a dynamical way to affect the gene expression time series. In light of the above, it is evident that network structure, physicochemical parameters and TF activity time courses can not all be extracted from a single approach.

Conclusion

Multiplex time series data (e.g. microarray, ChIP-on-chip and protein mass spectroscopy) holds a great promise for the construction of the network of cellular processes and the calibration of the many associated physical chemical parameters. We have demonstrated these concepts in the context of transcription regulation understood through the analysis of microarray time series data. Casting the approach in a probabilistic framework has allowed us to address the uncertainties in microarray data. Our approach was found to be robust to error in the microar-

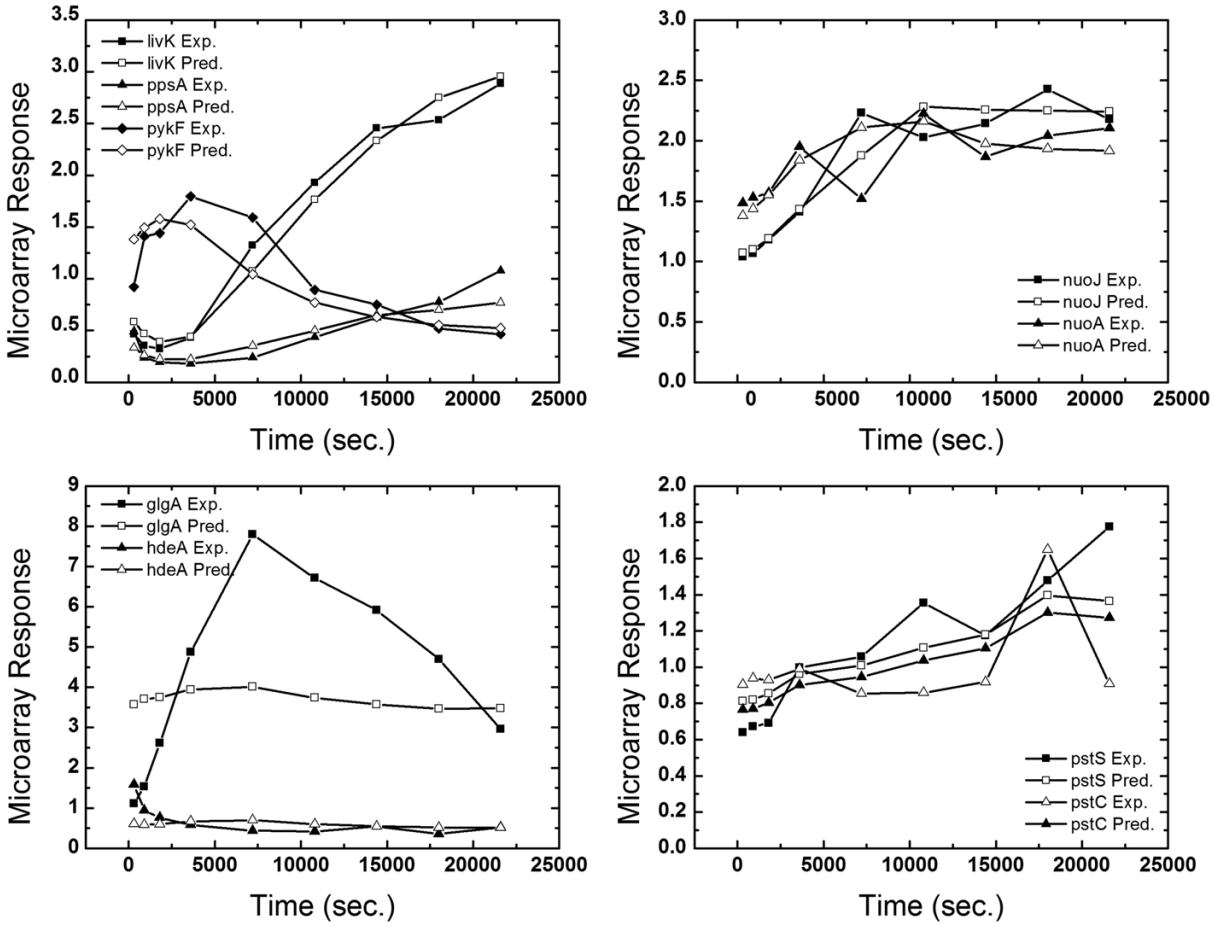

Figure 6
Comparison of predicted (hollow markers) and observed (solid markers) microarray response for a) nuoJ, nuoA, b) arcA, c) livK, ppsA, pykF, d) pstC and pstS.

ray data and mistakes in a proposed regulatory network. Our approach compliments other methods (e.g. gene ontology, phylogenetic and sequence analysis) when used as a part of a wider network discovery/quantification algorithm. Given its robustness, its capacity to refine and quantify complex networks of cellular processes, and the potential for extension to other multiplex bioanalytical data, we believe that our approach has great potential in the pure and applied life sciences.

Authors' contributions

AS, KT and PJO formulated the problem; and developed the theoretical model framework. AS and KT carried out the development, and implementation of the numerical algorithms. All authors participated in the writing of the manuscript, and have read and approved the manuscript.

Appendices

A. Numerical methods

The numerical methods used for simulating time evolution of mRNA populations, solving the calibration inverse problem by determining of TF time courses and model parameters are as follows. The latter parameters are sets of TF binding constants $\underline{\underline{Q}}$ and saturation limiting transcription rate coefficients $\underline{k}^{max}$ and mRNA degradation rate constants $\underline{\lambda}$.

Fast and accurate solution of the ODE model is crucial to construct thousands of gene expression levels to find the optimum model parameters in a practical time. For the *i*-th gene, mRNA population, R_i, time evolution is computed using an implicit Euler method

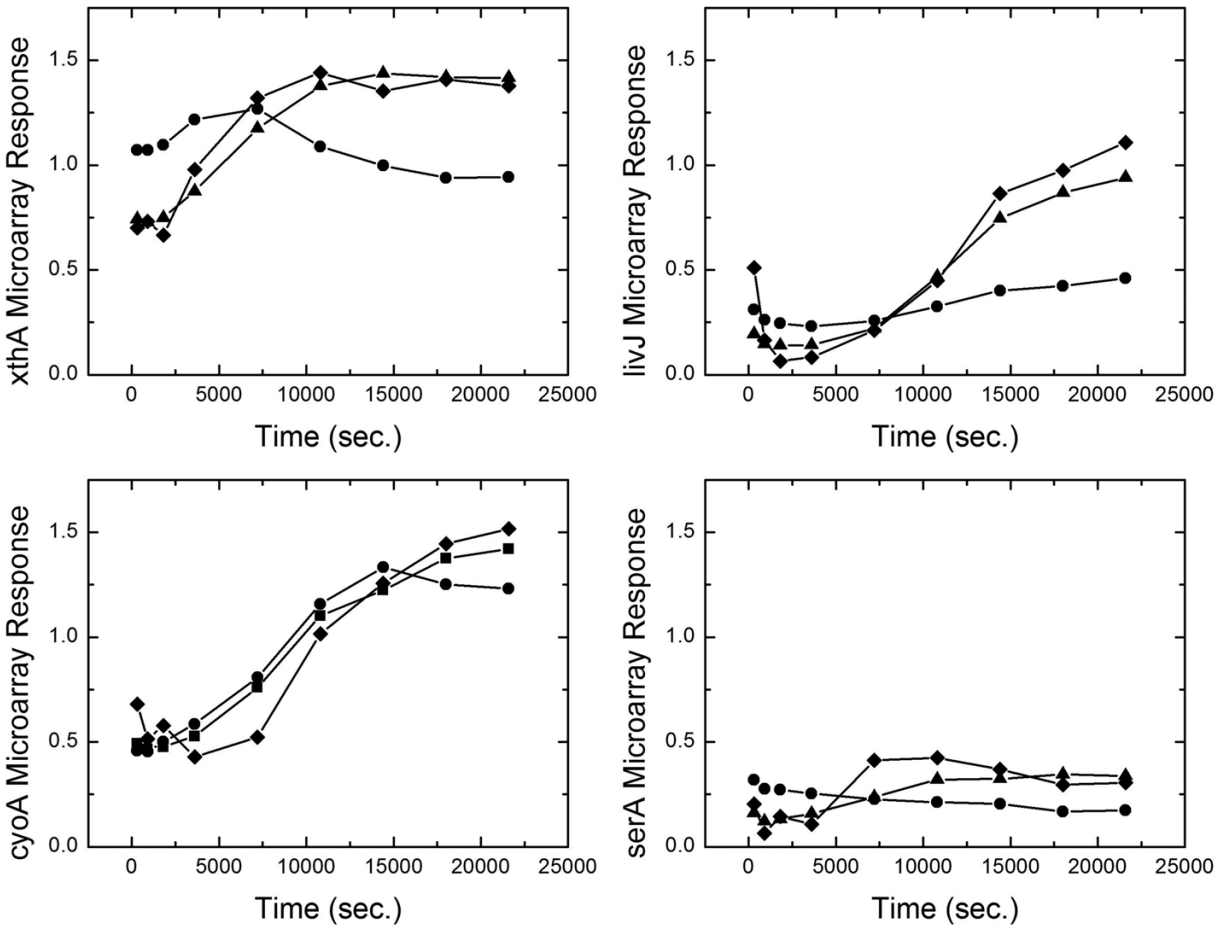

Figure 7
The predicted microarray response of xthA, livJ, cyoA and serA is enhanced after adding interactions suggested from promoter sequence analysis. Diamonds indicate the experimental microarray response; Circles indicate the predicted microarray response before adding the suggested interactions; Triangles indicate the predicted microarray response after adding the suggested interactions

$$R_i(t_{n+1}) = \frac{R_i(t_n) + \Delta t_{n+1} k_i(t_n)}{1+\lambda_i}. \qquad (A.1)$$

The time step Δt_{n+1} used is adaptive and depends on the maximum component of the rate vector $\underline{k}$. The microarray expression level at a given experimental time is predicted as the relative abundance of mRNA populations to their reference state at initial time

$$m_{i\ell} = m_i(t_\ell) = \frac{R_i(t_\ell)}{R_i(t_0)}, \quad \ell = 1, \cdots N_{micro}. \quad (A.2)$$

In solving $\partial\rho/\partial\underline{\Lambda} = 0$, a gradient steepest descent approach suffers from slow convergence. We overcome this via a combined steepest descent/simulated annealing approach. The key to efficiently solve the inverse problem cited above is to use an iterative alternating parameter approach. The calibration starts by minimization of the microarray error E^{cDNA} with respect to TF binding constants $\underline{\underline{Q}}$. To reduce the computational cost, we utilize the exact solution of (Eq. 6) [46],

$$R_0 m_i(t)\exp(\lambda_i t) - R_0 = \int_0^t k_i(t')\exp(\lambda_i t')dt'. \qquad (A.3)$$

The latter establishes an integral equation for $k_i(t)$. Solving for $k_i(t)$ at the given experimental microarray times yields a computationally efficient algebraic approach that allows the use of a simulated annealing algorithm [47] to find the optimum values for binding constants. The solution is achieved by discretizing the time profile of k_i over a grid of microarray experimental times and then interpolating it as a continuous piecewise linear function,

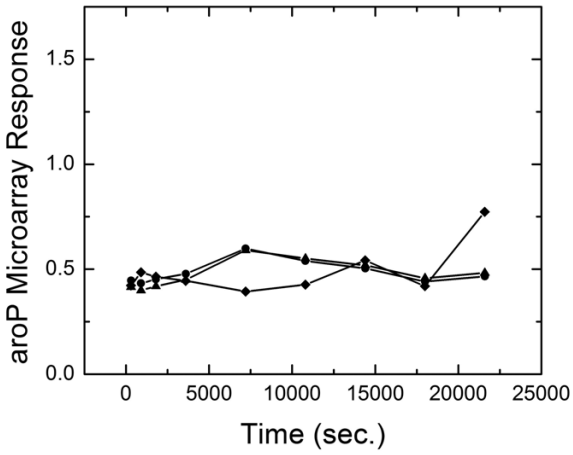

Figure 8
The predicted microarray response of aroP shows no significant improvement as we add the suggested CRP interaction. Diamonds indicate the experimental microarray response; Circles indicate the predicted microarray response before adding the suggested interactions; Triangles indicate the predicted microarray response after adding the suggested interactions.

$$k_i(t) = \frac{k_i^{l+1} - k_i^l}{t_{l+1} - t_l}(t - t_l) + k_i^l, \quad t_l < t < t_{l+1}, \quad l = 0, \cdots N_{micro}. \qquad (A.4)$$

With (A.4) we can evaluate the above integral analytically

$$R_0 m_{il}^{obs} \exp(\lambda_i t_l) - R_0 = \sum_{s=0}^{l-1} \left(\begin{array}{l} k_i^s \left\{ \frac{\exp(\lambda_i t_{s+1}) - \exp(\lambda_i t_s) + \lambda_i \exp(\lambda_i t_s)(t_{s+1} - t_s)}{\lambda_i^2 (t_{s+1} - t_s)} \right\} \\ + k_i^{s+1} \left\{ \frac{\exp(\lambda_i t_s) - \exp(\lambda_i t_{s+1}) + \lambda_i \exp(\lambda_i t_{s+1})(t_{s+1} - t_s)}{\lambda_i^2 (t_{s+1} - t_s)} \right\} \end{array} \right) \qquad (A.5)$$

Our lack of knowledge about the initial value of k_i (i.e. k_i^0), gives us one less equation than the number of unknowns. The simplest way to overcome this difficulty is via a linear extrapolation between the points at t_2 and t_1 to the point at t_0. Higher order extrapolations were tested and proven not to be very advantageous in this case, especially if we have frequent microarray measurements for early times. With this,

$$k_i^0 = \frac{k_i^2 - k_i^1}{t_2 - t_1}(t_0 - t_1) + k_i^1. \qquad (A.6)$$

(A.5) and (A.6) give us a linear system that can be solved for k_i^l, $l = 0, \cup N_{micro}$. We use the resulting k_i^l to construct a new error measure $\tilde{E}$,

$$\tilde{E}_i = \sum_{l=1}^{N_{micro}} \sum_{j=1}^{N_s^{(i)}} \left(\ln \left(\frac{\left(Q_{ij} T_{n_{ij}}(t_l) \right)^{\frac{1+b_{ij}}{2}}}{1 + Q_{ij} T_{n_{ij}}(t_l)} \right) - \ln k_i^l \right)^2, \qquad (A.7)$$

where n_{ij} is the type of TF that binds to the j-th site on gene i. We find that, if the rate integral equation is accurately solved, then minimizing this error is equivalent to minimizing the microarray error. Applying simulated annealing enhance the likelihood that we get as close as needed to the global minimum of $\tilde{E}_i$. When the resulting solution fails by increasing the error due to numerical instability, we switch to a steepest descent scheme.

For the TF activities we solve the discretized temporal regularization functional differential equations, $\partial\rho/\partial \underline{T} = 0$, with no flux boundary conditions [8] for its activity time course implicitly. For the j-th TF at the $(n+1)$-th iteration, one obtains

$$\frac{-\omega\Delta s}{\Delta t_f}\left(T_j^{n+1}(t_{l-1}) + T_j^{n+1}(t_{l+1})\right) + \left(1 + \frac{2\omega\Delta s}{\Delta t_f}\right) T_j^{n+1}(t_l) = -\Delta s \frac{\partial E}{\partial T_j^n(t_l)} + T_j^n(t_l), \qquad (A.8)$$

$$\left(1 + \frac{\omega\Delta s}{\Delta t_f}\right) T_j^{n+1}(t_1) - \frac{\omega\Delta s}{\Delta t_f} T_j^{n+1}(t_2) = -\Delta s \frac{\partial E}{\partial T_j^n(t_1)} + T_j^{\boldsymbol{n}}(t_1), \qquad (A.9)$$

and

$$\left(1 + \frac{\omega\Delta s}{\Delta t_f}\right) T_j^{n+1}(t_{N_{times}}) - \frac{\omega\Delta s}{\Delta t_f} T_j^{n+1}(t_{N_{times}-1}) = -\Delta s \frac{\partial E}{\partial T_j^n(t_{N_{times}})} + T_j^n(t_{N_{TF}}). \qquad (A.10)$$

Where $l = 1, \cup(N_{times} - 1)$, ω is the regularization coefficient and Δs is chosen small enough by line search to assure that E^{cDNA} is minimized. For the j-th set of equations, one must restrict the analysis to those genes regulated by that TF. The above linear system is efficiently solved using the Thomas algorithm for tridiagonal linear systems [48]. The remaining parameters (i.e. mRNA degradation rate coefficients $\underline{\lambda}$, and transcription limiting rate $\underline{k}^{max}$) are found by a steepest descent based on E^{cDNA}.

If only the microarray data was provided, and in absence of direct information and physical measurements on the binding constants and TF activities, it is clear that there is a degeneracy in the solution for this problem. This means that there are many states in the parameter space that have the similar E^{cDNA}. For example if Q_{ij}, $T_{n_{ij}}$ satisfy the error minimization criterion, then for all $\varepsilon > 0$ also $\tilde{Q}_{ij} = \varepsilon Q_{ij}$,

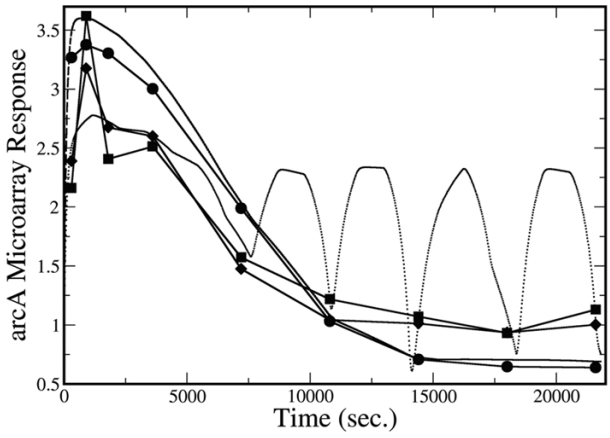

Figure 9
Comparison of predicted and observed microarray response for arcA. Diamonds indicate the experimental microarray response; squares indicate the experimental microarray response with 25% added noise; Circles indicate the predicted microarray response before adding 25% noise. Using data with 25% added noise, doted-line indicates the simulated microarray response when no regularization was imposed on the TF activity time courses, while the dashed-line is the microarray response when regularization was imposed on the TF activity time courses.

$\tilde{T}_{n_{ij}} = \frac{1}{\varepsilon} T_{n_{ij}}$ satisfy the same criterion. A normalization procedure is imposed on the solution at every step in the iterative inversion by assuming knowledge of the temporal average of each TF to be $\bar{T}_{n_{ij}}$ as follows

$$T^{norm}_{n_{ij}}(t) = \frac{T_{n_{ij}}(t) N_{micro} \bar{T}_{n_{ij}}}{\sum_{l=1}^{N_{micro}} T_{n_{ij}}(t_l)}, \qquad (A.11)$$

and

$$Q^{norm}_{ij} = \frac{Q_{ij} \sum_{l=1}^{N_{micro}} T_{n_{ij}}(t_l)}{N_{micro} \bar{T}_{n_{ij}}}. \qquad (A.12)$$

This is self-consistent since it eliminates the cited above *QT* degeneracy.

B. Symmetry rule and microarray inversion

For a general class of models used here, there is a TF up/down regulation symmetry that leads to a multiplicity in the determination of $\underline{T}(t)$. Notably there are $2^{N_{TF}}$ solutions of the microarray inversion problem that are equally viable unless some knowledge of $\underline{\underline{b}}$ is provided. This is proved for our model as follows. The control function H (Eq. 2) contains factors of the form $x^b/(1 + x)$ where b is $(b_{ij} + 1)/2$ and x is $Q_{ij}\ T_{n_{ij}}$. Note that $x/(1 + x) = 1/(1 + 1/x)$. Thus an up regulation with $Q_{ij}\ T_{n_{ij}}$ is equivalent to a down regulation with $1/(Q_{ij}\ T_{n_{ij}})$. This suggests that unless for each TF we know b_{ij} for at least one gene the inversion will allow two equally probable answers corresponding to $b_{ij} = \pm 1$ (either the correct result Q_{ij} and $T_{n_{ij}}$ for all i of the given TF type, or $1/T_{n_{ij}}$ with binding constant $1/Q_{ij}$). This implies that for each TF type n we must find at least one gene for which the nature of the regulation (up versus down) is known. This means that if $\underline{\underline{b}}$ is written in a sparse form as a N_g row by N_{TF} column matrix, then at least one entry in each column must be known. (see Table 1).

Acknowledgements

This project was supported by the United States Air Force (through the Defense Advanced Research Projects Agency), the United States Department of Energy (Genomics: Genome to Life Program; DE-FC02-02ER63446 and DE-FG02-05ER25676) and IBM Life Sciences Institute of Innovations as well as the College of Arts and Sciences and the Office of the Vice President for research (Indiana University) as general support for the Center for Cell and Virus Theory.

References

1. DeRisi JL, Iyer VR, Brown PO: **Exploring the Metabolic and Genetic Control of Gene Expression on a Genome Scale.** *Science* 1997, **278(5338):**680-686.
2. Sauter G, Simon R, Hillan K: **Tissue microarrays in drug discovery.** *Nature Reviews Drug Discovery* 2003, **2(12):**962-972.
3. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA microarray.** *Science* 1995, **270(5253):**467-470.
4. Gerhold D, Rushmore T, Caskey CT: **DNA chips: promising toys have become powerful tools.** *Trends in Biochemical Sciences* 1999, **24:**168-173.
5. Brown PO, Botstein D: **Exploring the new world of the genome with DNA microarray.** *Nature Genetics* 1999, **21:**33-37.
6. Chitler SV: **DNA microarrays:Tools for the 21(st) century.** *Combinatorial Chemistry and High throughput Screening* 2004, **7(6):**531-537.
7. Debouck C, Goodfellow PN: **DNA microarrays in drug discovery and development.** *Nature Genetics* 1999, **21:**48-50.
8. Sayyed-Ahmad A, Tuncay K, Ortoleva PJ: **Toward Automated Cell Model Development through Information Theory.** *Journal of Physical Chemistry A* 2003, **107(49):**10554-10565.
9. Rashevsky N: **Mathematical Biophysics Physico-Mathematical Foundations of Biology.** *Volume 1*. 3rd edition. New York , Dover Publications; 1960.
10. Slepchenko BM, Schaff JC, Macara I, Loew LM: **Quantitative cell biology with the virtual cell.** *Trends in Cell Biology* 2003, **13(11):**570-576.
11. Weitzke EL, Ortoleva PJ: **Simulating cellular dynamics through a coupled transcription, translation, metabolic model.** *Computational Biology and Chemistry* 2003, **27(4-5):**469-480.

12. Mendes P, Kell DB: **Nonlinear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation.** *Bioinformatics* 1998, **14(10):**869-883.
13. Nelson DE, Ihekwaba AEC, Elliott M, Johnson JR, Gibney CA, Foreman BE, Nelson G, See VH C. A., Spiller DG, Edwards SW, McDowell HP, Unitt JF, Sullivan E, Grimley R, Benson N, Broomhead D, Kell DB, White MRH: **Oscillations in NF-kappa B signaling control the dynamics of gene expression.** *Science* 2004, **306(5296):**704-708.
14. Shang YF, Hu X, DiRenzo J, Lazar MA, Brown M: **Cofactor dynamics and sufficiency in estrogen receptor-regulated transcription.** *Cell* 2000, **103(6):**843-852.
15. Wilson DL, Buckley MJ, Helliwell CA, Wilson IW: **New normalization methods for cDNA microarray data.** *Bioinformatics* 2002, **19(11):**1325-1332.
16. Hyduke DR, Rohlin L, Kao KC, Liao JC: **A software package for cDNA microarray data normalization and assessing confidence intervals.** *OMICS: A Journal of Integrative Biology* 2003, **7(3):**227-234.
17. Symth G, Speed T: **Normalization of cDNA Microarray Data.** *Methods* 2003, **31:**265-273.
18. Shmulevich I, Dougherty ER, Kim S, Zhang W: **Probabilistic Boolean Networks: a rule based uncertainty for gene regulatory networks.** *Bioinformatics* 2002, **18(2):**261-274.
19. Pe'er D, Regev A, Elidan G, Friedman N: **Inferring subnetworks from perturbed expression profiles.** *Bioinformatics* 2001, **17(Supplement):**S215-S224.
20. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data.** *Journal of Compuational Biology* 2000, **7(3/4):**601-620.
21. Li Y, Cambell C, Tipping M: **Bayesian automatic relevance determination algorithms for classifying gene expression data.** *Bioinformatics* 2002, **18(10):**1332-1339.
22. Azuaje F: **A cluster validity framework for genome expression data.** *Bioinformatics* 2002, **18:**319-320.
23. Bolshakova N, Azuaje F: **Cluster validation techniques for genome expression data.** *Signal Processing* 2003, **83:**825-833.
24. Karp PD, Riley M, Paley SM, Pellegrini-Toole A, Krummenacker M: **EcoCyc:Encyclopedia of Escherichia coli genes and metabolism.** *Nucleic Acids Research* 1998, **26(1):**50-53.
25. Liebermeister W: **Linear modes of gene expression determined by independent compnent analysis.** *Bioinformatics* 2002, **18(1):**51-60.
26. Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedroff NV: **Fundamental patterns underlying gene expression profiles: Simplicity from complexity.** *Proceedings of National Academy of Science* 2000, **97(15):**8409-8414.
27. Holter NS, Maritan A, Cieplak M, Fedroff NV, Banavar JR: **Dynamics modeling of gene expression data.** *Proceedings of National Academy of Science* 2001, **98(4):**1693-1698.
28. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP: **Network component analysis:Reconstruction of regulatory signals in biological systems.** *Proceedings of National Academy of Science* 2003, **100(26):**15522-15527.
29. Kao KC, Yang YL, Boscolo R, Sabatti C, Roychowdhury V, Liao JC: **Transcriptome-based determination of multiple transcription regular activities in Escherichia coli by using network component analysis.** *Proceedings of National Academy of Science* 2004, **101(2):**641-646.
30. Glanemann C, Loos A, Gorret N, Willis LB, O'Brien XM, Lessard PA, Sinskey AJ: **Disparity between Changes in mRNA Abundance and Enzyme Activity in Corynebacterium Gluamicum: Implications for DNA Microarray Analysis.** *Applied Microbiology and Biotechnology* 2003, **61(1):**61-68.
31. Gardner TS, Bernardo D, Lorenz D, Collins JJ: **Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling.** *Science* 2003, **301:**102-105.
32. Wu H, Su Z, Mao F, Olman V, Xu Y: **Prediction of functional modules based on comparative genome analysis and Gene Ontology application .** *Nucleic Acids Research* 2005, **33:**2822-2837.
33. Brown CT, Callan CG: **Evolutionary comparisons suggest many novel cAMP response protein binding sites in Escherichia coli.** *Proceedings of National Academy of Science* 2004, **101(8):**2404-2409.
34. Krauss G: **Biochemistry of Signal Transduction and Regulation.** Wiley-VCH Verlag; 2003.
35. Laidler KJ, Meiser JH: **Physical Chemistry.** 2nd edition. Boston , Houghton Mifflin Company; 1995.
36. Halford SE, Marko JF: **How Do Site-Specific DNA-Binding Proteins find their targets?** *Nucleic Acids Research* 2004, **32(10):**3040-3052.
37. Seiser C, Posch M, Thompson N, Kuhn LC: **Effect of transcription inhibitors on the iron-dependent degradation of transferrin receptor mRNA.** *Journal of Biological Chemistry* 1995, **270(49):**29400-29406.
38. Shannon CE: **A Mathematical Theory of Communication.** *Bell System Technical Journal* 1948:379-423,623-656.
39. Shannon CE, Weaver W: **The Mathematical Theory of Communication.** Urbana , University of Illinois Press; 1949.
40. Ortoleva P, Berry E, Brun Y, Fan J, Fontus M, Hubbard K, Jaqaman K, Jarymowycz L, Navid A, Sayyed-Ahmad A, Shreif Z, Stanley F, Tuncay. K, Weitzke E, WU LC: **The Karyote Physico-Chemical Genomic, Proteomic, Metabolic Cell Modeling System.** *OMICS: A Journal of Integrative Biology* 2003, **7(3):**269-283.
41. van Bakel H, Holstege FC: **In control: systematic assessment of microarray performance.** *EMBO reports* 2004, **5(10):**964-969.
42. Yue H, Eastman PS, Wang BB, Minor J, Doctolero MH, Nuttall RL, Stack R, Becker JW, Montgomery JR, Vainer M, Johnston R: **An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression.** *Nucleic Acids Research* 2001:e41-1.
43. Yuen T, Wurmbach E, Pfeffer RL, Ebersole BJ, Sealfon SC: **Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays.** *Nucleic Acids Research* 2002, **30(10):**e48.
44. Novak JP, Sladek R, Hudson TJ: **Characterization of variability in large-scale gene expression data: implications for study design.** *Genomics* 2002, **79(1):**104-113.
45. Salgado H, Santos-Zavaleta A, Gama-Castro S, Millen-Zarate D, Diaz-Peredo E, Sanchez-Solano F, Perez-Rueda E, Bonavides-Martinez C: **RegulonDB(version 3.2):transcriptional regulation and operon organization in Escherichia coli K-12.** *Nucleic Acids Research* 2001, **29(1):**72-74.
46. Hirsch MWS Stephen and Devaney, Robert L.: **Differential Equations, Dynamical Systems, & An Introduction to Chaos.** San Diego , Academic Press; 2004.
47. Corana A, Marchesi M, Martini C, Rdella S: **Minimizing Multimodal Functions Of Continuous-Variables with Simulated Annealing Algorithm.** *ACM Transactions on Mathematical Software* 1987, **13(3):**262-280.
48. Fletcher CAJ: **Computational Techniques for Fluid Dynamics.** Berlin , Springer-Verlag; 1988.