Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computers and Chemical Engineering 29 (2005) 565–576

Computers & Chemical Engineering

www.elsevier.com/locate/compchemeng

Large-scale inference of the transcriptional regulation of *Bacillus subtilis*

Anshuman Gupta[a], Jeffrey D. Varner[b], Costas D. Maranas[a,*]

[a] *Department of Chemical Engineering, The Pennsylvania State University, University Park, PA 16802, USA*
[b] *Process Science Group, Genencor International In, 925 Page Mill Rd, Palo Alto, CA 94304, USA*

Received 19 December 2003; received in revised form 22 March 2004
Available online 12 October 2004

Abstract

This paper addresses the inference of the transcriptional regulatory network of *Bacillus subtilis*. Two inference approaches, a linear, additive model and a non-linear power-law model, are used to analyze the expression of 747 genes from *B. subtilis* obtained using Affymetrix GeneChip® arrays under three different experimental conditions. A robustness analysis is introduced for identifying confidence levels for all inferred regulatory connections. Both the linear and non-linear methods produce candidate networks that share a scale-free or a “hub-and-spoke” topology with a small number of global regulator genes influencing the expression of a large number of target genes. The two computational approaches in tandem are able to identify known global regulators with a high level of confidence. The linear model is able to identify the interactions of highly expressed genes, particularly those involved in genetic information processing, energy metabolism and signal transduction. Conversely, the non-linear power-law approach tends to capture development regulation and specific carbon and nitrogen regulatory interactions.

Keywords: *Bacillus subtilis*; Transcriptional regulation; Signal transduction

1. Introduction

The data explosion currently overwhelming biology presents a challenging paradox: “you can see everything, but understand very little”. Sequencing a genome, which only a few years ago was a tremendous feat, is now routine. Many examples of transcript array analysis can be found in the literatures (Chu et al., 1998; Cohen, Pilpel, Mitra, & Church, 2002; Gardner, di Bernardo, Lorenz, & Collins, 2003; Lee, Rinaldi, & Robert, 2002; Spellman et al., 1998; Wen et al., 1998); an exponential increase since the ground breaking work of Brown and co-workers on *S. cerevisiae* (DeRisi, Iyer, & Brown, 1997). Genomics, proteomics and metabolomics are now the cutting edge of physiological analysis (Gill et al., 2002; Herrgard, Covert, & Palsson, 2003; Ideker et al., 2001; Misra et al., 2002; Oh, Rohlin, Kao, & Liao, 2002; Stephanopoulos, Hwang, Schmitt, Misra, & Stephanopoulos, 2002). These technologies when married with physiological measurements, genome-wide transcription measurements and genetic sequence generate huge tracts of integrated data. Buried inside this vast amount of interconnected information lay the output of millions of years of evolution. Gene expression, the process by which a genetic blueprint is turned into a working component of the organism through an intermediate mRNA message, is a meticulously managed endeavor. Understanding the control of gene expression has long been an important challenge because its understanding is central to our ability to steer an organism in directions other than for which it was programmed. Expression data alone does not hold the answer to this puzzle. However, if no transcript is present, there will be no protein (the reverse is not always true). Thus, understanding the regulation of gene expression networks is a necessary though not sufficient first step towards elucidating the flow of information in biological systems.

A number of studies have been conducted and frameworks proposed for the purpose of extracting regulatory networks from gene expression data. Most early network inference

[*] Corresponding author. Tel.: +1 814 863 9958; fax: +1 814 865 7846.
E-mail address: costas@psu.edu (C.D. Maranas).

0098-1354/$ – see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compchemeng.2004.08.030

methods relied primarily on clustering genes on the basis of their expression profiles (D'Haeseleer, Wen, Fuhrman, & Somogyi, 1999; Dougherty et al., 2002; Eisen, Spellman, Brown, & Botstein, 1998; Wen et al., 1998). Recently, there has been considerable interest in developing computational tools that go beyond answering the question of whether two or more genes have similar expression profiles. Instead, the central question that is being raised is whether we can uncover, hidden within gene expression data, the signature, extent and directionality of interactions between different genes. In other words, rather than simply grouping genes with similar expression profiles, new methods attempt to learn gene regulatory patterns from expression data. Broadly, these methods can be classified into two distinct categories based on their fundamental treatment of gene interactions. *Deterministic model-based* methods assume there exists a deterministic formalism $Y = f(X)$ that captures the effect of expression level of gene X on gene Y. Different choices for the function $f(\cdot)$ (e.g., linear, sigmoidal, etc.) give rise to many versions of model-based methods (D'Haeseleer et al., 1999; Gardner et al., 2003; Holter, Maritan, Cieplak, Fedoroff, & Banavar, 2001; Weaver, Workman, & Stormo, 1999; Zak, Gonye, & Schwaber, 2003). Conversely, *stochastic model-based* methods start by postulating that experimentally observed gene expression profiles correspond to samples drawn from an unknown multivariate probability distribution. Bayesian networks provide a popular alternative for achieving this objective by postulating a multivariate joint conditional probability model that explains the observed expression data (Friedman, Linial, Nachman, & Pe'er, 2000; Pe'er, Regev, Elidan, & Friedman, 2001).

Both deterministic and stochastic models have their respective advantages and disadvantages. The relative simplicity and computational tractability of deterministic models makes them amenable to inference of large-scale genome-wide transcriptional networks. However, these models are sensitive to over-fitting and prediction artifacts. Stochastic inference models such as Bayesian networks, because they take a probabilistic view of gene expression, are less sensitive to over-fitting. However, the application of Bayesian network models is limited due to their cumbersome treatment of cycles and their relatively large computational requirements. In the light of the advantages/limitations of the two approaches, a deterministic model-based approach is adopted to infer the underlying regulatory network of *Bacillus subtilis* on a global scale.

In addition to classifying gene network inference methods based upon the mathematical formalism used to model the regulation process, a further distinction can be made based upon how gene expression is handled within these formalisms. Boolean networks were among the first formalisms proposed to model gene interactions (Akutsu, Miyano, & Kuhara, 1999; Ideker, Thorsson, & Karp, 2000; Somogyi & Sniegoski, 1996). In this approach, genes are assumed to be either ON or OFF and the input–output relationships between them are modeled through deterministic logical functions (such as AND, OR, NOT, etc.). More recently, an extension of this approach to account for uncertainty in expression data has been proposed in the form of probabilistic Boolean networks (Akutsu, Miyano, & Kuhara, 2000; Shmulevich, Dougherty, Kim, & Zhang, 2002; Shmulevich, Lahdesmaki, Dougherty, Astola, & Zhang, 2003). However, in most real gene expression settings, Boolean idealizations may not be appropriate as genes are expressed at continuously varying intermediate expression levels (Jong, 2002). Consequently, more general approaches have been proposed which model mRNA expression level as a continuously varying quantity. These include linear weight modeling (D'Haeseleer et al., 1999; Weaver et al., 1999), ordinary differential equations (Chen, He, & Church, 1999) and S-systems (Akutsu et al., 2000; Kikuchi, Tominaga, Arita, Takahashi, & Tomita, 2003; Maki, Tominaga, Okamoto, Watanabe, & Eguchi, 2001; Savageau, 1998). In this work, we use a continuous description of gene expression since we do not observe any natural threshold values in our experimental data that can be used to unambiguously discretize the expression states.

2. Systems and methods

2.1. Scope of investigation

The regulatory network formed by 747 genes involved in the central metabolism of *B. subtilis* during fed-batch protease production is resolved from time series gene expression data collected from three different experimental conditions via two different deterministic model-based approaches. *B. subtilis* and related *Bacilli* are industrial workhorse organisms used for, among other things, protein production. *B. subtilis* is arguably the best characterized gram-positive bacteria. It has been sequenced (Kunst et al., 1997) and a large number of quantitative physiological studies of *B. subtilis* and related *Bacilli* are present in the literature (Christiansen, Christensen, & Nielsen, 2002; Christiansen, Michaelsen, Wumpelmann, & Nielsen, 2003; Dauner & Sauer, 2000, 2001; Dauner, Bailey, & Sauer, 2001a; Dauner, Storni, & Sauer, 2001b; Dauner et al., 2002; Sauer et al., 1997).

The three experimental conditions used for inference employ Affymetrix GeneChip$^{®}$ arrays for obtaining genome-wide expression data for: (i) a cradle-to-grave experiment (20 time points taken over the entire course of the fermentation); (ii) an amino acid pulse experiment (9 time points taken immediately after an amino acid pulse in mid-exponential growth); and (iii) an exponential growth phase experiment (5 time points taken during exponential growth). The use of Affymetrix arrays implies that our expression data are time-resolved absolute (non-condition scaled) transcript signals as opposed to the relative expression changes typically measured with cDNA arrays. The gene list includes genes involved in primary metabolism in addition to some known transcriptional regulators.

Two alternative model formulations, with different levels of computational complexity, are used to identify regulatory connections. The first approach, based upon modeling the dynamics of gene expression as a first-order, linear process (D'Haeseleer et al., 1999), assumes the expression level of a gene at a particular time-point is modeled as a linear combination of the concentration of all other genes at the previous time-point. The network connectivity, encoded by the coefficients in the linear combinations, is determined by minimizing the error between the predicted and experimental gene expression values. The second methodology generalizes the linear approach by postulating a non-linear differential equation model (Varner, 2000). In the non-linear framework a feedback law representing the control instructions governing expression is identified using tools from non-linear control theory. Putative connections between gene *j* and gene *k* are captured using a power-law expansion of the identified control function where the power-law exponents are determined directly from time-resolved gene expression data. A robustness analysis is carried out on the identified connections of both models to determine their level of confidence. The underlying principle of the robustness analysis is to assign a confidence level to each inferred connection by comparing the likelihood of inferring a regulatory coefficient of a given magnitude from real versus randomized data. We first explore the linear model followed by a complementary study using a non-linear model.

2.2. Linear model

The linear model considered here was first employed by (D'Haeseleer et al., 1999) and has been subsequently used by a number of other researchers (Alter, Brown, & Botstein, 2000; Gardner et al., 2003; Holter et al., 2000, 2001; Weaver et al., 1999; Yeung, Tegner, & Collins, 2002). In this approach, the rate of change of concentration of the mRNA species *i* is given by

$$\frac{\mathrm{d}z_i(t)}{\mathrm{d}t} = \sum_{j=1}^{N} w_{ij} z_j(t) \quad \forall i = 1, 2, \ldots, N \tag{2.1}$$

where $z_i(t)$ (pmol/g_{dw}) is the concentration of mRNA species *i* as measured at time-point *t* and w_{ij} is the regulatory coefficient capturing the regulatory effect of gene *j* on gene *i*. If $w_{ij} > 0$ then gene *j* up-regulates (activates) gene *i* while if $w_{ij} < 0$ then, gene *j* down-regulates (inhibits) the expression of gene *i*. If $w_{ij} = 0$, then no regulatory connection is implied between genes *i* and *j*. Given the discrete nature of the gene expression time series data, Eq. (2.1) can be approximated by the following set of linear algebraic equations:

$$\frac{z_i(t+1) - z_i(t)}{\Delta t} = \sum_{j=1}^{N} w_{ij} z_j(t) \quad \forall\, i = 1, 2, \ldots, N, \qquad t = 1, 2, \ldots, T-1 \tag{2.2}$$

Note that a discrete (forward) difference is employed for estimating the rate of change of expression (D'Haeseleer et al., 1999; Yeung et al., 2002). Alternatively, we also used an interpolation procedure for estimating the rate of change of gene expression. However, no significant differences in the results were observed over the simpler, forward difference approach. Eq. (2.2) is recast into matrix notation to yield

$$\dot{\mathbf{Z}}_{\mathbf{N}\times(\mathbf{T}-\mathbf{1})} = \mathbf{W}_{\mathbf{N}\times\mathbf{N}}\mathbf{Z}_{\mathbf{N}\times(\mathbf{T}-\mathbf{1})} \tag{2.3}$$

For most microarray time course experiments, the total number of genes investigated *N*, is much larger than the number of time-points *T*. This implies that the above system of equations is underdetermined because there are $N(T-1)$ equations and N^2 variables leading to multiple solutions. Singular value decomposition (SVD) is used to obtain the entire family of solutions that is consistent with the hypothesized linear model (Yeung et al., 2002). To this end, the transpose of Eq. (2.3) is taken in order to recast the system of equations in standard form (i.e., **Ax** = **b**):

$$(\mathbf{Z}^{\mathrm{T}})_{(\mathbf{T}-\mathbf{1})\times\mathbf{N}}(\mathbf{W}^{\mathrm{T}})_{\mathbf{N}\times\mathbf{N}} = (\mathbf{Z}^{\mathrm{T}})_{(\mathbf{T}-\mathbf{1})\times\mathbf{N}} \tag{2.4}$$

Subsequently, SVD is applied to Z^{T} yielding

$$(\mathbf{Z}^{\mathrm{T}})_{(\mathbf{T}-\mathbf{1})\times\mathbf{N}} = \mathbf{U}_{(\mathbf{T}-\mathbf{1})\times(\mathbf{T}-\mathbf{1})}\boldsymbol{\Sigma}_{(\mathbf{T}-\mathbf{1})\times\mathbf{N}}(\mathbf{V}^{\mathrm{T}})_{\mathbf{N}\times\mathbf{N}} \tag{2.5}$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix containing the $T-1$ non-zero singular values $\sigma_1, \sigma_2, \ldots, \sigma_{T-1}$ and **V** an orthogonal matrix containing the singular vectors $v_1, v_2, \ldots, v_N$ corresponding to *all* (both zero and non-zero) singular values. The particular solution for Eq. (2.4) that minimizes the L_2-norm is then given by

$$(\hat{\mathbf{W}}^{\mathrm{T}})_{\mathbf{N}\times\mathbf{N}} = \mathbf{V}_{\mathbf{N}\times\mathbf{N}}\boldsymbol{\Sigma}^{-1}_{\mathbf{N}\times(\mathbf{T}-\mathbf{1})}\mathbf{U}^{\mathrm{T}}_{(\mathbf{T}-\mathbf{1})\times(\mathbf{T}-\mathbf{1})}(\dot{\mathbf{Z}}^{\mathrm{T}})_{(\mathbf{T}-\mathbf{1})\times\mathbf{N}} \tag{2.6}$$

where $\boldsymbol{\Sigma}^{-1}$ is a diagonal matrix with values $\sigma_1^{-1}, \sigma_2^{-1}, \ldots, \sigma_{T-1}^{-1}$ (Yeung et al., 2002). The general solution for the underdetermined system of equations represented by Eq. (2.4) is given by

$$(\mathbf{W}^{\mathrm{T}})_{\mathbf{N}\times\mathbf{N}} = (\hat{\mathbf{W}}^{\mathrm{T}})_{\mathbf{N}\times\mathbf{N}} + \mathbf{C}_{\mathbf{N}\times(\mathbf{N}-\mathbf{T}+\mathbf{1})}(\hat{\mathbf{V}}^{\mathrm{T}})_{(\mathbf{N}-\mathbf{T}+\mathbf{1})\times\mathbf{N}} \tag{2.7}$$

where $\hat{\mathbf{V}}$ is the null-space matrix and **C** a matrix of arbitrary scalars (Yeung et al., 2002). All possible alternate network configurations that are consistent with the experimental data are embedded within Eq. (2.7). From this family, the sparsest network is determined by choosing the scalar coefficient matrix **C** such that the number of zero entries in $\mathbf{W}^{\mathrm{T}}$ is maximized. This is achieved by solving the Linear Programming

(LP) problem:

$$\text{minimize } \sum_{i,j}(w_{ij}^{+} + w_{ij}^{-})$$

$$\text{subject to } \hat{w}_{ij} + \sum_{k=1}^{N-T+1} c_{jk}\hat{v}_{ki} = w_{ij}^{+} - w_{ij}^{-} \quad \forall i = 1, 2, \ldots, N,$$

$$j = 1, 2, \ldots, N, \quad w_{ij}^{+} \geq 0, \quad w_{ij}^{-} \geq 0$$

$$\forall i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, N$$

where $\hat{w}_{ij}$ is the (*i*,*j*)th element of $\hat{\mathbf{W}}^{\mathrm{T}}$, $\hat{v}_{ki}$ the (*k*,*i*)th element of $\hat{\mathbf{V}}^{\mathrm{T}}$ and $w_{ij}^{+} + w_{ij}^{-}$ is the (*i*,*j*)th element of $\mathbf{W}^{\mathrm{T}}$. The right-hand side of the equality constraint quantifies the absolute deviation of the (*i*,*j*)th element of $\mathbf{W}^{\mathrm{T}}$ from zero. The objective function minimizes the total of all such deviations so that sparsity can be achieved. A key feature of the LP optimization model as formulated above is that it decomposes over the *j* index. This implies that instead of solving one large-scale optimization problem involving all genes, it can be solved sequentially for each gene. This decomposable structure of the problem can be exploited for (a) parallelizing the solution algorithm and (b) limiting the amount of computational resources expended if only a sub-network involving a sub-set of all genes needs to be inferred (Yeung et al., 2002).

Using basic LP principles, it can be shown that at the optimal solution of the LP model, each gene can be regulated by at most ($T-1$) genes. This is because for a given *j*, the total number of equations in the LP model is *N*, while the total number of variables is $2N + (N - T + 1)$. Since each basic feasible solution for the LP must have *N* basic variables, of which ($N - T + 1$) will be the c_{jk} variables (since they are free variables and are not being directly forced in any particular direction by the objective function), the remaining ($T - 1$) variables will be the absolute deviations. Note that for a given (*i*,*j*)th element, only one of the two deviation variables can be non-zero as otherwise the basis would have linearly dependent columns. Thus, a particular regulatory connection will be inferred exclusively as being activating or inhibiting.

2.3. Non-linear model

The mass balance equations governing the specific concentration of the *j*th mRNA species, denoted as $z_j(t)$ (pmol/g$_{\mathrm{dw}}$) is given by

$$\frac{\mathrm{d}z_j(t)}{\mathrm{d}t} = r_{T,z_j}(t)u_{z_j}(t) - (\hat{r}_{\mathrm{g}}(t) + \beta_{z_j})z_j(t),$$

$$y_{z_j}^{\mathrm{M}}(t) = f_{z_j}(\mathbf{z}(t), \mathbf{k}), \quad j = 1, 2, \ldots, N \tag{2.8}$$

where $r_{T,z_j}(t)$ denotes the maximum specific rate of expression of gene *j* (specific expression rate in the absence of control input), β_{z_j} the rate constant governing the specific degradation of transcript *j* (specific rate of degradation assumed to be first-order with transcript concentration) and $\hat{r}_{\mathrm{g}}(t)$ the specific growth rate. The maximum specific rate of transcription and the rate constants governing mRNA degradation are unknown and are assumed to be randomly distributed within physiologically possible ranges. These distributions are sampled through a Monte Carlo procedure (see Section 3). The specific growth rate is measured from fermentation data. The quantity $y_{z_j}^{\mathrm{M}}(t)$ denotes the raw array signal for transcript *j* at time *t* and $f_{z_j}(\mathbf{z}(t), \mathbf{k})$ represents the relationship between intracellular concentration and array signal for transcript *j*, where $\mathbf{k}$ denotes a vector of parameters contained in the function $f_{z_j}(\mathbf{z}(t), \mathbf{k})$.

The quantity $u_{z_j}(t)$ in Eq. (2.8) denotes the control input put forth by the organism to regulate the expression of gene *j*. If a mechanistic understanding of the regulation of gene *j* were known, it could be used to capture how the state of the organism affects the expression of gene *j*. Many examples of this approach exist in the literature for both deterministic as well as stochastic gene expression scenarios (Arkin, Ross, & McAdams, 1998; Bailey et al., 1983; Lee & Bailey, 1984a, 1984b; McAdams & Arkin, 1997, 1998, 1999; Rao & Arkin, 2001; Rao, Wolf, & Arkin, 2002; Wolf & Arkin, 2002; Wong, Gladney, & Keasling, 1997). However, we assume that we do not have the control mechanism and as such are not able to build a mechanistic-based representation of $u_{z_j}(t)$. Rather, we use the transcript profiling data to identify a feedback control law $u_{z_j}^{\mathrm{M}}(t)$ that is guaranteed to produce a model estimated array signal that tracks the measured value (Csete & Doyle, 2002).

Define the error between the experimental array signal for transcript *j* at time *t* ($y_{z_j}^{\mathrm{E}}(t)$) and the model predicted array signal for transcript *j* at time *t* $y_{z_j}^{\mathrm{M}}(t)$ as

$$\varepsilon_j(t) \equiv y_{z_j}^{\mathrm{M}}(t) - y_{z_j}^{\mathrm{E}}(t) \tag{2.9}$$

We propose the prediction error be governed by the linear dynamics:

$$\frac{\mathrm{d}\varepsilon_j}{\mathrm{d}t} = -\lambda_j \varepsilon_j(t) \quad \forall j \tag{2.10}$$

where $\lambda_j \geq 0$ for every *j*. Differentiating Eq. (2.9) and substituting Eqs. (2.8) and (2.10) yields

$$\sum_{q=1}^{N} \frac{\partial f_j}{\partial z_q}[r_{T,z_q}(t)u_{z_q}(t) - (\hat{r}_{\mathrm{g}}(t) + \beta_{z_q})z_q(t)] = \frac{\mathrm{d}y_j^{\mathrm{E}}}{\mathrm{d}t} - \lambda_j \varepsilon_j(t) \quad \forall j \tag{2.11}$$

Eq. (2.11) describes the relationship between the control input governing the expression of gene *j* and the error between the estimated and measured array signals.

It is possible to solve for $u_{z_j}(t)$ as a function of the prediction error if the relationship between the array signal and mRNA concentration were known for each gene, i.e., the function $f_{z_j}(\mathbf{z}(t), \mathbf{k})$. The exact relationship relating array signal to concentration is not known. However, in all cases in vitro transcripts consisting of a cocktail of five different mRNA species of known concentration ranging from 0.5 to 160 pM were added to all samples directly before the analysis. This internal standard (referred to as IVTs) provides a *very*

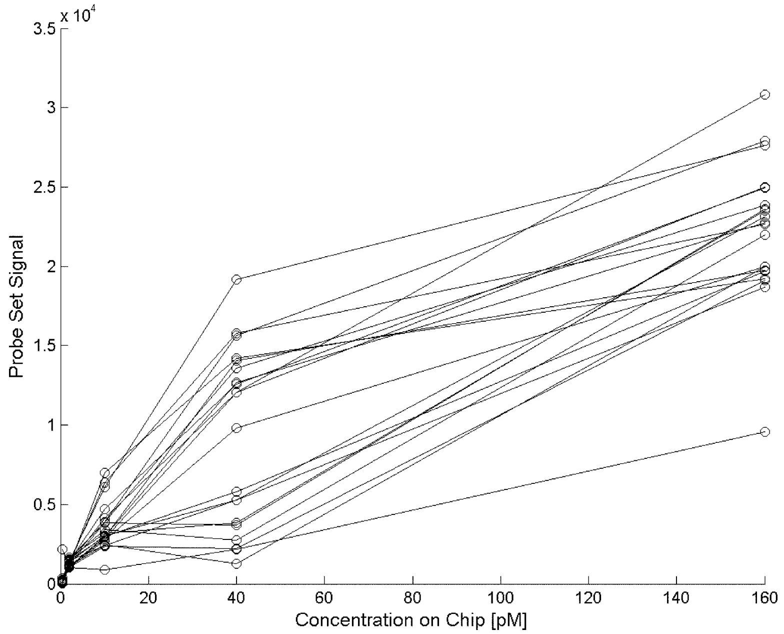

Fig. 1. IVT probe set signals vs. chip concentration for the cradle-to-grave experiment. The open circles denote a probe set signal for a given concentration at each time point. For example, at 40 pM, a single transcript species was measured at each time point (total of 20 measurements for the cradle-to-grave run). A correction factor was introduced to convert chip concentration to intracellular concentration.

approximate means of determining the relationship between chip concentration and array signal (once we have an estimate of chip concentration it is possible to back calculate an approximate physiological concentration). As there is considerable variance amongst the 5 IVT species at a single concentration (see Fig. 1), we use a Monte Carlo approach to estimate a set of possible $f_{z_j}(\mathbf{z}(t), \mathbf{k})$'s. We approximate $f_{z_j}(\mathbf{z}(t), \mathbf{k})$ as a set of piece-wise linear functions:

$$y^{\mathrm{M}}_{z_j}(t) = m(\theta z_j(t)) + b,$$

$$(m, b) = \begin{cases} (m_1, b_1), & 0 \le y^{\mathrm{E}}_{z_j}(t) \le S_1 \\ \vdots & \vdots \\ (m_n, b_n), & S_{n-1} < y^{\mathrm{E}}_{z_j}(t) \le S_n \end{cases} \qquad (2.12)$$

where the slope and y-intercept (m_j,b_j) are determined by fitting a line between sequential IVT concentration pairs and their corresponding signals (one line per signal region per iteration) and θ denotes the conversion between physiological and chip concentration. Given Eq. (2.12) we can solve Eq. (2.11) for $u_{z_q}(t)$

$$u^{\mathrm{M}}_{z_j}(t) = \frac{1}{r_{T,z_j}(t)}\left\{\delta_{z_j}(t)\left(\frac{y^{\mathrm{M}}_j(t) - b}{m\theta}\right) + \frac{1}{m\theta} \times \left(\frac{y^{\mathrm{E}}_j(t+\Delta t) - y^{\mathrm{E}}_j(t-\Delta t)}{2\Delta t} - \lambda(y^{\mathrm{M}}_j(t) - y^{\mathrm{E}}_j(t))\right)\right\} \qquad (2.13)$$

where $\delta_{z_j}(t)$ is defined as

$$\delta_{z_j}(t) \equiv (\hat{r}_{\mathrm{g}}(t) + \beta_{z_j}) \qquad (2.14)$$

A central difference is used to approximate the array signal derivative using interpolated array data (step size is Δt = 0.1 h). Eq. (2.13) represents the feedback input that is guaranteed to produce a simulated array signal that will converge to the experimental signal as $t \to \infty$. Thus, it approximates the control input used by the organism to produce the transcriptional program captured in the experiment.

To probe the connectivity of the expression network, we expand $u^{\mathrm{M}}_{z_j}(t)$ in terms of the estimated specific mRNA concentration vector $\mathbf{z}$. In contrast with the linear model that looks for interactions that lie on or near a hyperplane in expression space, the non-linear model utilizes the power-law expansion:

$$u_{z_j}(t) \equiv \frac{1}{z^{\max}} \prod_{q=1}^{N} z_q^{\gamma_{jq}}(t) \qquad (2.15)$$

which represents a curvilinear hypersurface. As was true in the linear case, the expansion exponent γ_{jq} denotes the sensitivity of the control of expression of gene j to the scaled concentration of transcript q, where the scaling factor $z^{\max}$ denotes the maximum estimated concentration determined over all expressed species and time. Taking the natural log of Eq. (2.15) yields the linear function:

$$\hat{u}_{z_j}(t) = \eta + \sum_{q=1}^{N} \gamma_{jq} \hat{z}_q(t) \qquad (2.16)$$

where $\hat{}$ denotes the natural log of the corresponding quantity in Eq. (2.15). Eq. (2.16) represents a family of planes in the log transformed control-transcript space where the sensitivity coefficients γ_{jq} and intercept η can be determined using multivariate regression.

Note that the networks inferred from both the linear and non-linear models are correlational in nature and not causal. This stems from the fact that the various regulatory relationships as inferred by the two models are determined purely by minimizing the error between the predicted expression profiles (under the particular model definition) and the experimentally obtained expression profiles. No mechanistic detail is taken into account in terms of whether a particular regulatory interaction is biologically feasible at the transcriptional level. In the spirit of previous works on inferring gene networks from microarray data (Alter et al., 2000; D'Haeseleer et al., 1999; Holter et al., 2000, 2001; Tegner, Yeung, Hasty, & Collins, 2003; Weaver et al., 1999; Yeung et al., 2002), our goal is to efficiently infer networks on a global scale and generate a "rough draft" of the network topology, using which more detailed and local analysis can be conducted in the future.

2.4. Permutation-based significance analysis

Both modeling approaches, by minimizing the error between the predicted gene expression value and the experimental value, generate a list of putative gene–gene regulatory interactions. However, it is unclear which of the inferred connections are real and which are simply artifacts of the inference process. For instance, it can be shown for the linear model, that each gene will be regulated by at most ($T - 1$). Clearly, this property arises from the way sparsity is enforced through the LP model and there is no biological justification for it. Such model bias is typically addressed by enforcing threshold (cut-off) values whereby connections for which the absolute value of the inferred regulatory coefficient is higher than some pre-specified threshold are considered significant. The choice of the threshold value is usually ad hoc and symmetric in the sense that the same threshold is enforced on both activating and inhibiting regulatory connections. To address this issue, we calculate a confidence estimate on the inferred connections based upon the probability that the connection will arise in random expression data, more exactly, row-column permuted data (Good, 2000; Pesarin, 2001). Thus, the confidence of a connection between genes i and j with an inferred regulatory coefficient value of w^*, denoted as $c_{ij}(w^*)$, is defined as

$$c_{ij}(w^*) = \frac{N^{\mathrm{a}}(w_{ij} \geq w^*)}{N^{\mathrm{a}}(w_{ij} \geq w^*) + N^{\mathrm{r}}(w_{ij} \geq w^*)} \tag{2.17}$$

where $N^{\mathrm{a}}(w_{ij} \geq w^*)$, $N^{\mathrm{r}}(w_{ij} \geq w^*)$ denote the number of regulatory connections inferred from the actual/randomized data with a regulatory coefficient value greater than w^* (see Fig. 2). For instance, if for a given value of w^*, no connections are inferred from the randomized data, then a confidence level of 100% is assigned to all connections inferred from the actual data with a regulatory coefficient higher than w^*. The inferred transcriptional network would thus be characterized by regulatory connections having different confidence levels. The confidence level metric can be used as an additional degree of freedom for studying the topological properties of the inferred network. Increasing the confidence level and analyzing the regulatory connections that survive yields hypotheses regarding the biological processes being captured in the expression data on the time-scale of the sampling frequency.

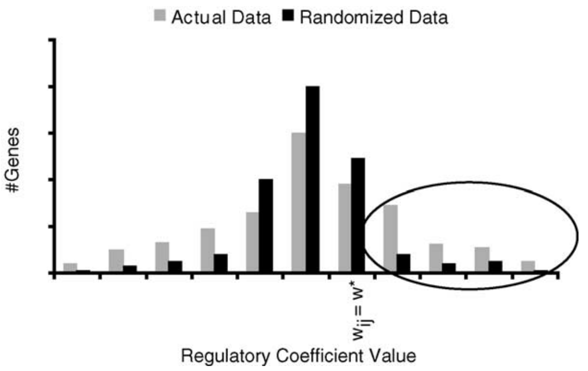

Fig. 2. The distributions obtained for the actual and randomized data sets are superimposed on top of each other. The confidence level for a given regulatory coefficient value w^* is then determined by computing the ratio of the number of inferred connections from the actual data that have a regulatory coefficient value greater than w^* to the total (sum of actual and randomized) number of regulatory connections with value greater than w^*.

3. Discussion

The network topology inferred for the three different experimental data sets using the linear model is shown in Fig. 3. These plots correspond to the sparsest regulatory networks. Rows corresponding to genes with no regulatory effect have been eliminated. The most striking feature of the inferred networks is the existence of regulatory bands, a characteristic, which indicates the existence of a small number of global regulators or "hubs" that influence a large number of other genes. Such "hub-and-spoke" topologies (also referred to as scale-free) have been observed for metabolic networks (Fell & Wagner, 2000; Jeong, Tombor, Albert, Oltvai, & Barabasi, 2000) and protein–protein interaction networks (Jeong, Mason, Barabasi, & Oltvai, 2001). Two key features of such networks are (i) their robustness to the random failure of the nodes of the network and (ii) relatively short paths between any two nodes in the network. Figs. 4 and 5 highlight the results of the permutation-based significance analysis for the linear and non-linear model. Regulatory coefficients with higher absolute value are inferred with higher confidence as shown in Fig. 4. Fig. 5 indicates the number of arcs that are inferred for a given confidence level. As expected, the fraction of inferred connections that are preserved is inversely proportional to confidence level. However, even though fewer number of arcs survive strict tolerance level requirement, the underlying scale-free topology of the net-

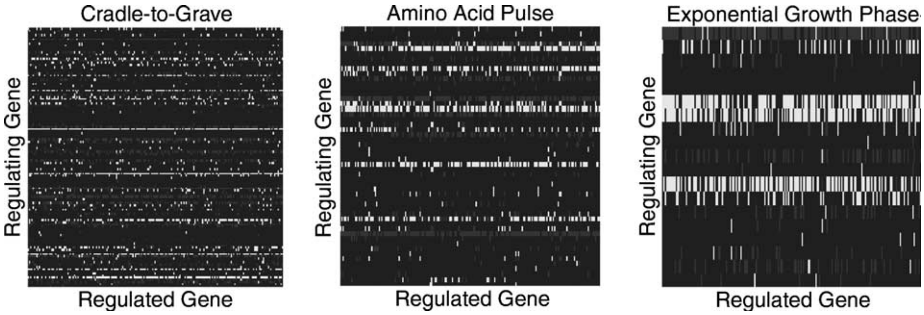

Fig. 3. Network topology inferred for the three experiments using the linear model. White signifies activation or positive regulation whereas grey denotes inhibition or negative regulation. Note that rows corresponding to genes not regulating any other genes are eliminated for clarity of presentation.

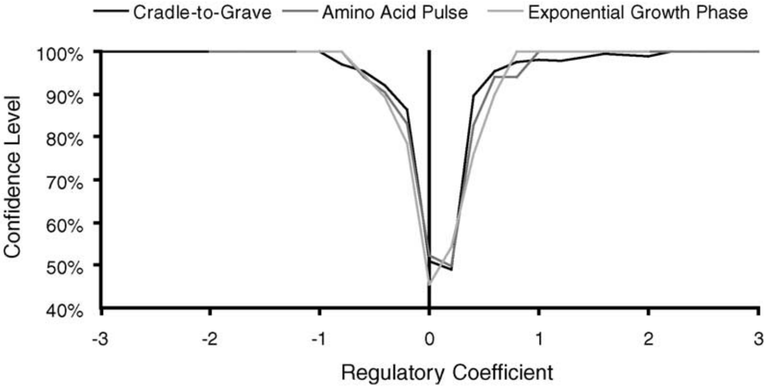

Fig. 4. Permutation-based significance analysis results for the linear model. For a given confidence level, asymmetric thresholds are obtained for activation and inhibition. For instance, for a minimum confidence level of 90% for the cradle-to-grave experiment, the activation and inhibition regulation thresholds are 0.416 and −0.324, respectively. Such thresholds are expected to be different for different experimental conditions as shown in the figure. However, the inhibition regulation threshold is seen to be smaller in absolute value than the activation regulation threshold for all three experiments.

work is still preserved as shown in Fig. 6. Also, the scale-free topology that is observed in the real expression data is not preserved when the data is randomized. The degree distributions for the randomized data sets for all three experimental conditions are closer to uniform distributions rather than exponential distributions implying that the observed scale-free topologies are not artifacts of the proposed inference procedure.

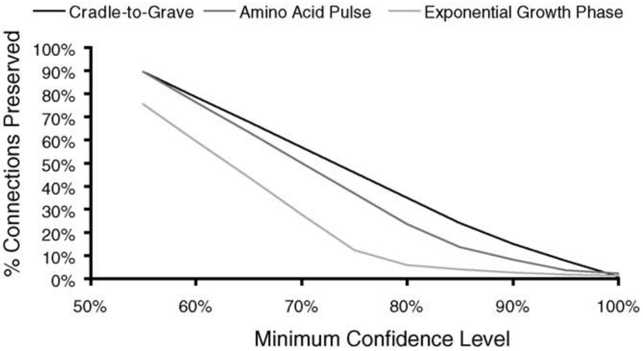

Fig. 5. Fraction of regulatory connections preserved as minimum confidence level is increased. A relatively small fraction of the total inferred connections (15% for cradle-to-grave, 8% for amino acid pulse and 3% for exponential growth phase) survive when a high level of confidence (>90%) is imposed.

Table 1 lists some of the global regulators identified by the linear model from the cradle-to-grave study. All the regulator genes identified in the base case regulatory network are retained even at higher confidence levels though with reduced out-degrees. A number of genes involved in amino acid metabolism are identified as regulator genes. These include *gbsA*, *argG*, *ytcF*, *trpE*, *rocF*, *ysiB* and *cysK*. In addition, genes participating in carbohydrate metabolism, in particular glycolysis/gluconeogenesis and the citrate cycle (TCA) are also inferred as regulatory genes. The glycolysis genes include *pdhC*, *gap*, and *acoL* while the TCA cycle genes include *sucC*, *citB*, *sdhA*, *odhB* and *pckA*. A number of genes that are involved in more than one functional category are also uncovered by the linear model. For instance, the *gap* gene is involved in both glycolysis and amino acid metabolism while *acoL* is involved in the TCA cycle as well as amino acid metabolism in addition to glycolysis. Energy metabolism genes involved in oxidative phosphorylation such as *ctaE*, *atpB* and *sdhA* are also inferred as regulator genes. Various genetic information processing genes are also inferred as regulator genes. These include *secE* (protein export, sorting

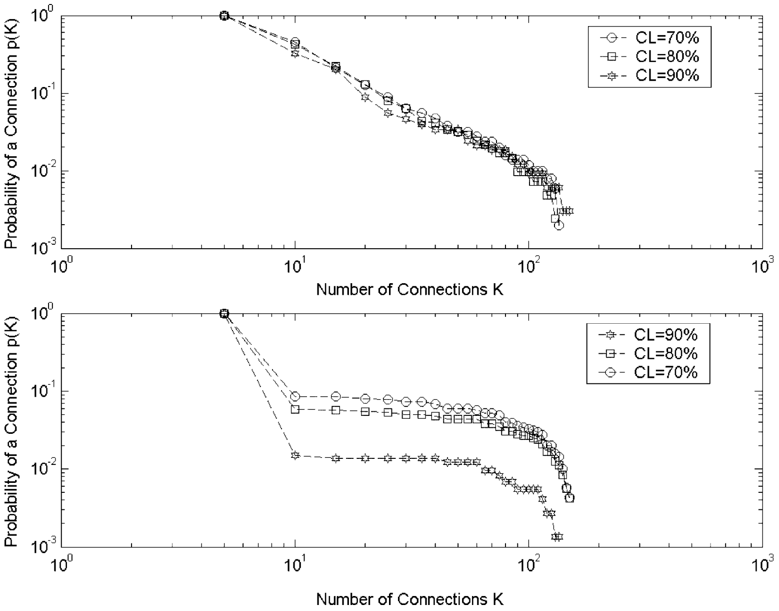

Fig. 6. Number of connections K vs. the normalized probability of at least K connections for the network inferred from the cradle-to-grave expression time series using the linear model (top panel) and non-linear model (bottom panel). The parameter K is the total-degree (sum of in- and out-degrees) of a particular node and $p(K)$ is the probability of finding a node with at least K connections (total-degree). The linear inference appears approximately linear on a log–log scale for all confidence levels suggesting a power-law relationship that is characteristic of scale-free networks. The non-linear distribution exhibits agreement with a power-law distribution for <10 connections after which the probability plateaus and then decays for a large number of connections. The probability of finding a connection in the non-linear regime is almost always lower than the sparse linear model, with the majority of genes having less than 10 connections.

Table 1
Identified global regulators inferred from the cradle-to-grave data set by the linear model

Gene	Function
secE	Preprotein translocase SecE subunit
phrE	Regulator of the activity of phosphatase rapE
ctaE	Cytochrome caa3 oxidase subunit III
gbsA	Glycine betaine aldehyde dehydrogenase
sucC	Succinyl-CoA synthetase beta chain
argG	Argininosuccinate synthase
atpB	ATP synthase a chain
pdhC	Pyruvate dehydrogenase
sigW	RNA polymerase ECF-type sigma factor
ytcF	S-adenosylmethionine decarboxylase
gapA	Glyceraldehyde 3-phosphate dehydrogenase
trpE	Anthranilate synthase component 1
citB	Aconitate hydratase
sdhA	Succinate dehydrogenase flavoprotein subunit
csn	Chitosanase
rocF	Arginase
odhB	2-Oxoglutarate dehydrogenase
pckA	Phosphoenolpyruvate carboxykinase
hpr	Transcriptional regulator for sporulation initiation
acoL	Acetoin dehydrogenase E3 component
rbsK	Ribokinase
ysiB	Enoyl-CoA hydratase
phrA	Phosphatase rapA inhibitor
ald	Stage V sporulation protein N
cysK	Cysteine synthetase A
rapA	Response regulator aspartate phosphatase
sigH	Sporulation-specific sigma factor

and degradation), *sigW* (RNA polymerase sigma factor), *hpr* (transcriptional regulator for peptide transport and sporulation initiation) and *ald* (stage V sporulation protein N).

Comparison of the networks inferred for the three experimental conditions identified several consensus regulatory relationships. Specifically, the two genes *yloH* and *phrA* are identified as a global activator and inhibitor, respectively, from all three datasets by the linear model. *yloH* is a key component of the transcription machinery as it encodes for the ω-subunit of the RNA polymerase. This gene is found to activate 22 other genes belonging to a wide range of functional categories including amino acid metabolism, carbohydrate and complex lipid metabolism and oxidative phosphorylation. In addition, this gene is found to up-regulate other genetic information processing genes, particularly those coding for aminoacyl-tRNA synthetases (*serS*, *gltX* and *thrS*) that are required for translation and DNA polymerase subunits (*dnaE* and *yorL*) that are required for DNA replication and repair. The *phrA* gene encodes for a 44 amino acid signaling protein involved in extracellular signaling that is required for timing the cell’s decision to choose a particular physiological state such as growth or sporulation (Jiang, Grau, & Perego, 2000; McQuade, Comella, & Grossman, 2001; Phillips & Strauch, 2002). This gene is found to down-regulate 18 other genes, of which 4 genes (*kinA*, *sigF*, *spo0A* and *spo0F*) are key participants in the initiation of sporulation (Stragier & Losick, 1996). These consensus results indicate that even though the

Table 2
Identified global regulators of known function inferred from the cradle-to-grave data set by the non-linear model

Gene	Function
spo0J	Stage 0 sporulation protein
spo0B	Sporulation initiation phosphotransferase
spo0F	Two-component response regulator involved in the initiation of sporulation
spoIIID	Transcriptional regulator of sigma-E and sigma-K dependent genes
ynzD	Hypothetical protein similar to *spo0E*
tnrA	Transcriptional pleiotropic regulator involved in global nitrogen regulation
phoR/*phoP*	Two-component sensor histidine kinase. Potential cognate response regulator is PhoP
cggR	Repressor of *gapA* expression
ccpA	Transcriptional regulator mediating carbon catabolite repression (LacI family)
kina	Two-component sensor histidine kinase involved in the initiation of sporulation
kinC	Two-component sensor histidine kinase involved in the initiation of sporulation
sigY	RNA polymerase ECF-type sigma factor
rapG(*F*,*K*,*D*,*I*)	Response regulator aspartate phosphatase
gerE	Transcriptional regulator required for the expression of late spore coat genes
lexA	Transcriptional repressor of the SOS regulon
glnR	Transcriptional repressor of the glutamine synthetase gene
oppC(*D*,*F*,*A*)	Oligopeptide transport system proteins
hprP	P-Ser-HPr phosphatase
sinR	Transcriptional regulator of post-exponential-phase responses genes
ald	Alanine dehydrogenase (stage V sporulation protein N)
secE(*F*)	Preprotein translocase subunit

linear model takes a relatively simplistic view of gene regulation, it can indeed uncover biologically relevant regulatory relationships.

The network topology recovered from the cradle-to-grave experiment using the non-linear model as a function of confidence level is shown in Fig. 6. As in the linear case, the inferred network is banded (indicative of a "hub-and-spoke" architecture). The non-linear inference produces a network in which the bulk of genes have less than 10 connections. There is approximately a 10% chance of finding a gene with more than 10 connections. The chance of finding a gene with a total degree of >100 is small (see Fig. 6). The non-linear model is able to capture regulatory interactions involved in the initiation of sporulation as well as global nitrogen and carbon metabolism. Some prominent regulatory genes inferred from the cradle-to-grave experiment via the non-linear model are listed in Table 2. Several known sporulation control genes, *spo0J*, *spo0B* and *kinC* are inferred as regulators with high confidence (Sonenshein, 2000). In addition to sporulation control, several potential nutritional regulators are also identified such as *tnrA*, a transcriptional pleiotropic regulator involved in global nitrogen metabolism (Beier, Nygaard, Jarmer, & Saxild, 2002; Brandenburg et al., 2002; Ferson, Wray, & Fisher, 1996; Fisher, 1999; Fisher, Brandenburg, & Wray, 2002; Fisher & Wray, 2002; Robichon et al., 2000; Wray, Ferson, & Fisher, 1997; Wray, Ferson, Rohrer, & Fisher, 1996; Wray, Zalieckas, Ferson, & Fisher, 1998; Wray, Zalieckas, & Fisher, 2001) and *ahrC*, a transcriptional regulator involved in the metabolism of arginine (Czaplewski, North, Smith, Baumberg, & Stockley, 1992; Dennis, Glykos, Parsons, & Phillips 2002; Holtham et al., 1999; Klingel, Miller, North, Stockley, & Baumberg, 1995; Miller, Baumberg, & Stockley, 1997; Stockley et al., 1998).

The non-linear model is able to capture known regulatory interactions at a confidence level above random. Consider the nitrogen metabolism regulators *tnrA* and *ahrC*. The non-linear model estimates that *tnrA* expression is connected with 61 genes (out-degree) at a confidence level of 60%. Of the 61, 20 genes are involved in nitrogen metabolism/amino acid biosynthesis or peptidoglycan biosynthesis, 7 are other probable regulatory genes (*rapG*, *rapF*, *spo0E*, *oppC*, *lexA*, *comP* and *comX*) and the remainder are carbon metabolism genes, for example, *glcK*, *gapB*, PTS system genes or DNA polymerases and ribosomal genes. Of the connections to nitrogen/amino acid biosynthetic genes, three predicted connections are known (*ureB*, *gltA* and *glnA* (Fisher, 1999; Belitsky, Wray, Fisher, Bohannon, & Sonenshein, 2000; Wray et al., 1996, 1997, 1998)) with a fourth gene, *gabD*, putatively linked to *gabP*, a permease downstream of *gabD*, known to be very strongly regulated by *tnrA* (Ferson et al.,

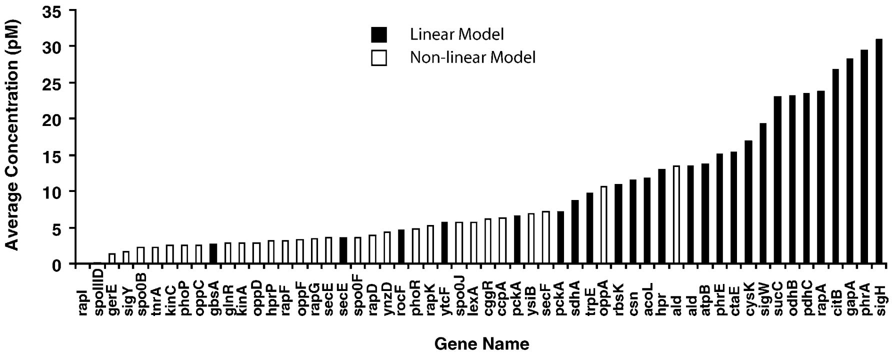

Fig. 7. Average concentrations of the regulator genes inferred by the two models.

1996; Wray et al., 1996, 1998). Consistent with the literature, *ahrC* is predicted to act as both an activator and repressor of arginine metabolism but also appears to be involved with the regulation of the metabolism of other amino acids (Czaplewski et al., 1992; Dennis et al., 2002; Holtham et al., 1999; Klingel et al., 1995; Miller et al., 1997; Stockley et al., 1998). *ahrC* is predicted to up-regulate *argD*, *pheA*, *hisB* and *gltA* and down-regulate *argG*, *tyrA*, *hisF*, *cysE*, *hisH*, *glyQ* and *trpA*.

Both models arrive at the same putative banded or "hub-and-spoke" architecture for the gene expression network of *B. subtilis*. However, the genes identified as hubs in the network by the respective models are quite different. The linear model seems much better able to capture interactions that occur among highly expressed genes (see Fig. 7). Moreover, the linear model is better able to capture self-regulatory interactions. By contrast, the non-linear model is better able to capture developmental regulation as well as specific carbon and nitrogen regulatory interactions, most of which take place by genes expressed at much lower levels. Hence, there seems to be gene expression regions where each inference approach performs best.

4. Conclusions

In this work, the large-scale inference of the transcriptional regulatory network of *B. subtilis* was addressed using two alternative computational methodologies; a linear, additive model and a non-linear, power-law model. The data on which the two inference techniques were applied consisted of a selected set of 747 genes whose expression was tracked over time using Affymetrix GeneChip® arrays under three different experimental conditions. The three time series data sets with 5, 9 and 20 time points, spanned a wide range of sampling frequencies.

The linear model extracted network connectivity by approximating the gene expression dynamics with a linear system of ordinary differential equations (ODEs). Discretization of this system of ODEs at the experimentally sampled time points resulted in an underdetermined system of linear equations. A two-step procedure was employed for solving this system of equations. First, singular value decomposition (SVD) was performed on the gene expression matrix in order to construct a general formalism satisfied by all possible networks that were consistent with both the experimental data and the chosen linear model. Subsequently, a sparsity hypothesis was enforced through a linear programming (LP)-based model for obtaining a particular network configuration. In contrast, a power-law model was used to explicitly track the non-linearities and their evolution through time in the *B. subtilis* transcriptional system. In addition to modeling non-linearities, the power-law methodology accounted for variability in the kinetic parameters and the relationship between array signal and transcript concentration by employing a Monte Carlo procedure. The control input of the organism was decomposed into the transcriptional program through the application of systems-theoretic tools and multivariate regression.

A robustness analysis was introduced for assigning confidence levels to all inferred regulatory connections. The underlying idea of this analysis was that by randomizing the expression data and then using the scrambled data in the inference procedure, any underlying model bias could be detected and eliminated. This bias elimination was achieved by imposing systematic, as opposed to arbitrary, cut-offs on the values of the inferred regulatory interactions.

Both inference methodologies were shown to result in transcriptional networks that exhibited scale-free, "hub-and-spoke" topologies. This corresponded to the existence of a relatively small number of global regulator genes that regulated the expression of a large number of target genes. The scale-free topology was found to be preserved even when very high confidence level requirements were imposed. The two modeling approaches were identified to be complementary with respect to their applicability in different gene expression regimes. Specifically, the linear model was able to identify interactions between highly expressed genes while the non-linear model was able to resolve the interactions between low expression genes. This observation highlights the fact that a number of alternative inference methodologies should be used in tandem for uncovering the wide spectrum of regulatory interactions that can be expected to exist at various concentration/temporal scales. In terms of future work, we are currently investigating the reason(s) for the applicability of the two models in the two distinct concentration regimes by systematically probing the two model formalisms using artificially generated data. Preliminary results indicate that the non-linear model is able to magnify the expression level of lowly expressed genes to a larger extent than the linear model. Detailed computational and theoretical analyses of the two models will form the focus of a future manuscript.

References

Akutsu, T., Miyano, S., & Kuhara, S. (1999). Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Proceedings of the Pacific Symposium on Biocomputing (PSB 1999)*, *4*, 17.

Akutsu, T., Miyano, S., & Kuhara, S. (2000). Inferring qualitative regulations in genetic networks and metabolic pathways. *Bioinformatics*, *16*, 727.

Alter, O., Brown, P. O., & Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *PNAS*, *97*, 10101.

Arkin, A., Ross, J., & McAdams, H. H. (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics*, *149*, 1633–1648.

Bailey, J. E., Hjortso, M., Lee, S. B., & Srienc, F. (1983). Kinetics of product formation and plasmid segregation in recombinant microbial populations. *Annals of New York Academic Science*, *413*, 71–87.

Beier, L., Nygaard, P., Jarmer, H., & Saxild, H. H. (2002). Transcription analysis of the *Bacillus subtilis* PucR regulon and identification of a *cis*-acting sequence required for PucR-regulated expression of genes involved in purine catabolism. *Journal of Bacteriology*, *184*, 3232–3241.

Belitsky, B. R., Wray, L. V., Jr., Fisher, S. H., Bohannon, D. E., & Sonenshein, A. L. (2000). Role of TnrA in nitrogen source-dependent repression of *Bacillus subtilis* glutamate synthase gene expression. *Journal of Bacteriology*, *182*, 5939–5947.

Brandenburg, J. L., Wray, L. V., Jr., Beier, L., Jarmer, H., Saxild, H. H., & Fisher, S. H. (2002). Roles of PucR, GlnR, and TnrA in regulating expression of the *Bacillus subtilis* ure P3 promoter. *Journal of Bacteriology*, *184*, 6060–6064.

Chen, T., He, H. L., & Church, G. M. (1999). Modeling gene expression with differential equations. *Proceedings of the Pacific Symposium on Biocomputing (PSB 1999)*, *4*, 29.

Christiansen, T., Christensen, B., & Nielsen, J. (2002). Metabolic network analysis of *Bacillus clausii* on minimal and semirich medium using (13)C-labeled glucose. *Metabolic Engineering*, *4*, 159–169.

Christiansen, T., Michaelsen, S., Wumpelmann, M., & Nielsen, J. (2003). Production of savinase and population viability of *Bacillus clausii* during high-cell-density fed-batch cultivations. *Biotechnology and Bioengineering*, *83*, 344–352.

Chu, S., DeRisi, J., Eisen, M. B., Mulholland, J., Botstein, D., Brown, P. O., & Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science*, *282*, 699.

Cohen, B. A., Pilpel, Y., Mitra, R. D., & Church, G. M. (2002). Discrimination between paralogs using microarray analysis: Application to the Yap1p and Yap2p transcriptional networks. *Molecular Biology of the Cell*, *13*, 1608.

Csete, M. E., & Doyle, J. C. (2002). Reverse engineering of biological complexity. *Science*, *295*, 1664–1669.

Czaplewski, L. G., North, A. K., Smith, M. C., Baumberg, S., & Stockley, P. G. (1992). Purification and initial characterization of AhrC: The regulator of arginine metabolism genes in *Bacillus subtilis*. *Molecular Microbiology*, *6*, 267–275.

Dauner, M., Bailey, J. E., & Sauer, U. (2001). Metabolic flux analysis with a comprehensive isotopomer model in *Bacillus subtilis*. *Biotechnology and Bioengineering*, *76*, 144–156.

Dauner, M., & Sauer, U. (2000). GC–MS analysis of amino acids rapidly provides rich information for isotopomer balancing. *Biotechnology Progress*, *16*, 642–649.

Dauner, M., & Sauer, U. (2001). Stoichiometric growth model for riboflavin-producing *Bacillus subtilis*. *Biotechnology and Bioengineering*, *76*, 132–143.

Dauner, M., Sonderegger, M., Hochuli, M., Szyperski, T., Wuthrich, K., Hohmann, H. P., Sauer, U., & Bailey, J. E. (2002). Intracellular carbon fluxes in riboflavin-producing *Bacillus subtilis* during growth on two-carbon substrate mixtures. *Applied and Environmental Microbiology*, *68*, 1760–1771.

Dauner, M., Storni, T., & Sauer, U. (2001). *Bacillus subtilis* metabolism and energetics in carbon-limited and excess-carbon chemostat culture. *Journal of Bacteriology*, *183*, 7308–7317.

Dennis, C. C., Glykos, N. M., Parsons, M. R., & Phillips, S. E. (2002). The structure of AhrC, the arginine repressor/activator protein from *Bacillus subtilis*. *Acta Crystallographica Section D: Biological Crystallography*, *58*, 421–430.

DeRisi, J. L., Iyer, V. R., & Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, *278*, 680.

D'Haeseleer, P., Wen, X., Fuhrman, S., & Somogyi, R. (1999). Linear modeling of mRNA expression levels during CNS development and injury. *Proceedings of the Pacific Symposium on Biocomputing (PSB 1999)*, *4*, 41.

Dougherty, E. R., Barrera, J., Brun, M., Kim, S., Cesar, R. M., Chen, Y., Bittner, M., & Trent, J. M. (2002). Inference from clustering with applications to gene-expression microarrays. *Journal of Computational Biology*, *9*, 105–126.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, *95*, 14863.

Fell, D. A., & Wagner, A. (2000). The small world of metabolism. *Nature Biotechnology*, *18*, 1121.

Ferson, A. E., Wray, L. V., Jr., & Fisher, S. H. (1996). Expression of the *Bacillus subtilis* gabP gene is regulated independently in response to nitrogen and amino acid availability. *Molecular Microbiology*, *22*, 693–701.

Fisher, S. H. (1999). Regulation of nitrogen metabolism in *Bacillus subtilis*: vive la difference!. *Molecular Microbiology*, *32*, 223–232.

Fisher, S. H., Brandenburg, J. L., & Wray, L. V., Jr. (2002). Mutations in *Bacillus subtilis* glutamine synthetase that block its interaction with transcription factor TnrA. *Molecular Microbiology*, *45*, 627–635.

Fisher, S. H., & Wray, L. V., Jr. (2002). Mutations in the *Bacillus subtilis* glnRA operon that cause nitrogen source-dependent defects in regulation of TnrA activity. *Journal of Bacteriology*, *184*, 4636–4639.

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, *7*, 601.

Gardner, T. S., di Bernardo, D., Lorenz, D., & Collins, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, *301*, 102–105.

Gill, R. T., Katsoulakis, E., Schmitt, W., Taroncher-Oldenburg, G., Misra, J., & Stephanopoulos, G. (2002). Genome-wide dynamic transcriptional profiling of the light-to-dark transition in *Synechocystis* sp. strain PCC 6803. *Journal of Bacteriology*, *184*, 3671–3681.

Good, P. I. (2000). *Permutation tests: A practical guide to resampling methods for testing hypotheses*. New York, NY: Springer Verlag.

Herrgard, M. J., Covert, M. W., & Palsson, B. O. (2003). Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Research*, *13*, 2423–2434.

Holter, N. S., Maritan, A., Cieplak, M., Fedoroff, N. V., & Banavar, J. R. (2001). Dynamic modeling of gene expression data. *PNAS*, *98*, 1693.

Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R., & Fedoroff, N. V. (2000). Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *PNAS*, *97*, 8409.

Holtham, C. A., Jumel, K., Miller, C. M., Harding, S. E., Baumberg, S., & Stockley, P. G. (1999). Probing activation of the prokaryotic arginine transcriptional regulator using chimeric proteins. *Journal of Molecular Biology*, *289*, 707–727.

Ideker, T. E., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., & Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, *292*, 929.

Ideker, T. E., Thorsson, V., & Karp, R. M. (2000). Discovery of regulatory interactions through perturbation: Inference and experimental design. *Proceedings of the Pacific Symposium on Biocomputing (PSB 2000)*, *5*, 302.

Jeong, H., Mason, S. P., Barabasi, A. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, *411*, 41.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, *407*, 651.

Jiang, M., Grau, R., & Perego, M. (2000). Differential processing of propeptide inhibitors of Rap phosphatases in *Bacillus subtilis*. *Journal of Bacteriology*, *182*, 303.

Jong, H. D. (2002). Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, *9*, 67.

Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K., & Tomita, M. (2003). Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics*, *19*, 643–650.

Klingel, U., Miller, C. M., North, A. K., Stockley, P. G., & Baumberg, S. (1995). A binding site for activation by the *Bacillus subtilis* AhrC protein, a repressor/activator of arginine metabolism. *Molecular & General Genetics*, *248*, 329–340.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S. C., Bron, S., Brouillet, S., Bruschi, C. V., Caldwell, B., Capuano, V., Carter, N. M., Choi, S. K., Codani, J. J., Connerton, I. F., Danchin, A., et al. (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature, 390,* 249–256.

Lee, S. B., & Bailey, J. E. (1984a). A mathematical model for lambda dv plasmid replication: Analysis of copy number mutants. *Plasmid, 11,* 166–177.

Lee, S. B., & Bailey, J. E. (1984b). A mathematical model for lambda dv plasmid replication: Analysis of wild-type plasmid. *Plasmid, 11*, 151–165.

Lee, T. I., Rinaldi, N. J., & Robert, F. (2002). Transcriptional regulatory network in *Saccharomyces cerevisiae*. *Science, 298*, 799.

Maki, Y., Tominaga, D., Okamoto, M., Watanabe, S., & Eguchi, Y. (2001). Development of a system for the inference of large scale genetic networks. *Proceedings of the Pacific Symposium on Biocomputing (PSB 2001), 6,* 446.

McAdams, H. H., & Arkin, A. (1997). Stochastic mechanisms in gene expression. *Proceedings of the National Academic Science USA, 94,* 814–819.

McAdams, H. H., & Arkin, A. (1998). Simulation of prokaryotic genetic circuits. *Annual Review of Biophysics and Biomolecular Structure, 27,* 199–224.

McAdams, H. H., & Arkin, A. (1999). It's a noisy business! Genetic regulation at the nanomolar scale. *Trends in Genetics, 15*, 65–69.

McQuade, R. S., Comella, N., & Grossman, A. D. (2001). Control of a family of phosphatase regulatory genes (phr) by the alternate sigma factor Sigma-H of *Bacillus subtilis*. *Journal of Bacteriology, 183,* 4905–4909.

Miller, C. M., Baumberg, S., & Stockley, P. G. (1997). Operator interactions by the *Bacillus subtilis* arginine repressor/activator, AhrC: Novel positioning and DNA-mediated assembly of a transcriptional activator at catabolic sites. *Molecular Microbiology, 26*, 37–48.

Misra, J., Schmitt, W., Hwang, D., Hsiao, L., Gullans, S., Stephanopoulos, G., & Stephanopoulos, G. (2002). Interactive exploration of microarray gene expression patterns in a reduced dimensional space. *Genome Research, 12*, 1112–1120.

Oh, M., Rohlin, L., Kao, K. C., & Liao, J. C. (2002). Global expression profiling of acetate-grown *Escherichia coli*. *The Journal of Biological Chemistry, 277*, 13175–13183.

Pe'er, D., Regev, A., Elidan, G., & Friedman, N. (2001). Inferring subnetworks from perturbed expression profiles. *Bioinformatics, 17*(S1), 215.

Pesarin, F. (2001). *Multivariate permutation tests with applications in biostatistics*. New York, NY: Wiley.

Phillips, Z. E. V., & Strauch, M. A. (2002). *Bacillus subtilis* sporulation and stationary phase gene expression. *Cellular and Molecular Life Sciences, 59*, 392–402.

Rao, C. V., & Arkin, A. P. (2001). Control motifs for intracellular regulatory networks. *Annual Review of Biomedical Engineering, 3*, 391–419.

Rao, C. V., Wolf, D. M., & Arkin, A. P. (2002). Control, exploitation and tolerance of intracellular noise. *Nature, 420*, 231–237.

Robichon, D., Arnaud, M., Gardan, R., Pragai, Z., O'Reilly, M., Rapoport, G., & Debarbouille, M. (2000). Expression of a new operon from *Bacillus subtilis*, ykzB–ykoL, under the control of the TnrA and PhoP–phoR global regulators. *Journal of Bacteriology, 182*, 1226–1231.

Sauer, U., Hatzimanikatis, V., Bailey, J. E., Hochuli, M., Szyperski, T., & Wuthrich, K. (1997). Metabolic fluxes in riboflavin-producing *Bacillus subtilis*. *Nature Biotechnology, 15,* 448–452.

Savageau, M. A. (1998). Rules for the evolution of gene circuitary. *Proceedings of the Pacific Symposium on Biocomputing (PSB 1998), 3,* 54.

Shmulevich, I., Dougherty, E. R., Kim, S., & Zhang, W. (2002). Probabilistic Boolean networks: A rule based uncertainty model for gene regulatory networks. *Bioinformatics, 18*, 261.

Shmulevich, I., Lahdesmaki, H., Dougherty, E. R., Astola, J., & Zhang, W. (2003). The role of certain post classes in Boolean network models of genetic networks. *PNAS, 100,* 10734–10739.

Somogyi, R., & Sniegoski, C. A. (1996). Modeling the complexity of genetic networks: Understanding multigenic and pleitropic regulation. *Complexity, 1*, 45.

Sonenshein, A. L. (2000). Control of sporulation initiation in *Bacillus subtilis*. *Current Opinion in Microbiology, 3*, 561–566.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell, 9*, 3273.

Stephanopoulos, G., Hwang, D., Schmitt, W., Misra, J., & Stephanopoulos, G. (2002). Mapping physiological states from microarray expression measurements. *Bioinformatics, 18*, 1054–1063.

Stockley, P. G., Baron, A. J., Wild, C. M., Parsons, I. D., Miller, C. M., Holtham, C. A., & Baumberg, S. (1998). Dissecting the molecular details of prokaryotic transcriptional control by surface plasmon resonance: The methionine and arginine repressor proteins. *Biosensors & Bioelectronics, 13*, 637–650.

Stragier, P., & Losick, R. (1996). Molecular genetics of sporulation in *Bacillus subtilis*. *Annual Review of Genetics, 30*, 297–341.

Tegner, J., Yeung, M. K. S., Hasty, J., & Collins, J. J. (2003). Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. *PNAS, 100*, 5944–5949.

Varner, J. D. (2000). Large-scale prediction of phenotype: Concept. *Biotechnology and Bioengineering, 69*, 664–678.

Weaver, D. C., Workman, C. T., & Stormo, G. D. (1999). Modeling regulatory networks with weight matrices. *Proceedings of the Pacific Symposium on Biocomputing (PSB 1999), 4*, 112.

Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Baker, J. L., & Somogyi, R. (1998). Large-scale temporal gene expression mapping of central nervous system development. *PNAS, 95*, 334.

Wolf, D. M., & Arkin, A. P. (2002). Fifteen minutes of fim: Control of type 1 pili expression in *E. coli*. *Omics, 6*, 91–114.

Wong, P., Gladney, S., & Keasling, J. D. (1997). Mathematical model of the lac operon: Inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose. *Biotechnology Progress, 13*, 132–143.

Wray, L. V., Jr., Ferson, A. E., & Fisher, S. H. (1997). Expression of the *Bacillus subtilis* ureABC operon is controlled by multiple regulatory factors including CodY, GlnR, TnrA, and Spo0H. *Journal of Bacteriology, 179*, 5494–5501.

Wray, L. V., Jr., Ferson, A. E., Rohrer, K., & Fisher, S. H. (1996). TnrA, a transcription factor required for global nitrogen regulation in *Bacillus subtilis*. *Proceedings of the National Academic Science of USA, 93*, 8841–8845.

Wray, L. V., Jr., Zalieckas, J. M., Ferson, A. E., & Fisher, S. H. (1998). Mutational analysis of the TnrA-binding sites in the *Bacillus subtilis* nrgAB and gabP promoter regions. *Journal of Bacteriology, 180*, 2943–2949.

Wray, L. V., Jr., Zalieckas, J. M., & Fisher, S. H. (2001). *Bacillus subtilis* glutamine synthetase controls gene expression through a protein–protein interaction with transcription factor TnrA. *Cell, 107*, 427–435.

Yeung, M. K. S., Tegner, J., & Collins, J. J. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *PNAS, 99*, 6163.

Zak, D. E., Gonye, G. E., Schwaber, J. S., & Doyle, F. J., III. (2003). Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: Insights from an identifiability analysis of an *in silico* network. *Genome Research, 13*, 2396–2405.