# *Predicting gene regulation by sigma factors in Bacillus subtilis from genome-wide data*

*Michiel J.L. de Hoon*[*] [1], *Y. Makita*[1], *S. Imoto*[1], *K. Kobayashi*[2], *N. Ogasawara*[2], *K. Nakai*[1] *and S. Miyano*[1]

[1]*Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan and* [2]*Graduate School of Biological Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara, 630-0101, Japan*

## ABSTRACT

**Motivation:** Sigma factors regulate the expression of genes in *Bacillus subtilis* at the transcriptional level. First we assess the ability of currently available gene regulatory network models to accurately infer gene regulation by sigma factors from gene expression data. Secondly, we consider improving the prediction accuracy by combining gene expression data with sequence information. Finally, we apply the resulting joint predictor to discover currently unknown gene regulations by sigma factors in *Bacillus subtilis*.

**Methods:** We determine the accuracy of sigma factor prediction from gene expression data using a fold-change analysis, Bayesian networks, dynamic models, and supervised learning based on coregulation. We show that the recently proposed method of combining a coregulation-based prediction with sequence information by summing the log-likelihood scores (Segal *et al.*, 2003), at least in our case, effectively ignores sequence information. We propose to use logistic regression to achieve a better balance between sequence and gene expression information.

**Results:** We show that the supervised learning method based on coregulation yields the most accurate prediction of sigma factors from gene expression data. We demonstrate in a leave-one-out experiment that the logistic regression model effectively combines gene expression data and sequence information. In a genome-wide search, highly significant logistic regression scores were found for several genes whose transcriptional regulation is currently unknown, allowing us to identify with high confidence the sigma factors regulating these genes. We provide the corresponding RNA polymerase binding sites to enable a straightforward experimental verification of our predictions.

**Keywords:** Gene regulation, Bayesian network, fold-change analysis, sigma factors, *Bacillus subtilis*

**Contact:** Email: mdehoon@ims.u-tokyo.ac.jp;
Telephone: +81-3-54495615; Fax: +81-3-54495442

## INTRODUCTION

The development of cDNA microarray technology has provided a huge amount of gene expression data. The methodology for analyzing such data is still in development. Recently, systems biology approaches have become increasingly popular, where the gene regulatory network and the interaction between genes are of prime interest.

Gene regulatory relations can be studied in gene disruptant experiments, in which the expression levels of all genes are measured after the expression of a transcription factor has been disrupted. A fold-change analysis is then performed to identify genes that are significantly up- or down-regulated due to the disruption, which may indicate that those genes are regulated by the transcription factor.

In time-course gene expression experiments, the expression levels of all genes are measured as a function of time following some perturbation in the environment of the organism. Dynamic models of gene regulation, such as differential equation models (Chen *et al.*, 1999) and dynamic Bayesian networks (Ong *et al.*, 2002; Kim *et al.*, 2003), take the time-dependence of the measurements into account by describing the gene expression levels at each time point in terms of the gene expression levels at the previous time point.

Alternatively, Bayesian networks inferred from cDNA microarray data have been proposed as a model of gene regulation (Friedman *et al.*, 2000; Imoto *et al.*, 2002a,b; Pe'er *et al.*, 2001). A Bayesian network shows how the expression level of each gene depends conditionally on a small set of parent genes.

[*] To whom correspondence should be addressed

Bayesian networks can be inferred from a set of static (time-independent) gene expression measurements of cell cultures acclimated to different environmental conditions, from gene disruptant experiments, as well as from time-course experiments, albeit without taking the time-dependence into account.

A fourth approach of inferring gene regulatory relations from expression data is based on coregulation. As genes regulated by the same transcription factor are likely to have similar gene expression patterns, unsupervised learning in the form of clustering gene expression data allows us to find coregulated genes (Segal *et al.*, 2003). In general, such an analysis will not reveal the corresponding transcription factor. However, when we are interested in finding additional genes regulated by a known transcription factor, coregulation can be applied as a supervised learning approach. Here, we compare the expression profile of a new gene to the expression profiles of the genes known to be regulated by that transcription factor.

Clustering gene expression data is often followed by searching for sequence motifs in the upstream region of coregulated genes. Segal *et al.* (2003) recently proposed to combine gene expression data and sequence information in a single Bayesian score function; Tamada *et al.* (2003) proposed a similar method in the framework of Bayesian networks. In this work, we found that in practice Segal's method may lead to an overestimation of the predictive power of gene expression data, to a degree that the sequence motif information is effectively ignored. To find a better balance between sequence information and gene expression data, we propose to use a logistic regression model to combine the two data sources.

Whereas the algorithmic aspects of these methods to infer gene regulatory relations have been well studied in the past, it is still unknown if these methods of inferring gene regulatory networks yield biologically correct results. Previous biological support of these methods has been limited to finding one or a few examples where the predicted regulatory relations agreed with biologically known results. To be able to predict currently unknown gene regulatory relations, however, this does not suffice, as we cannot known beforehand which of the large number of predicted gene regulatory relations in an inferred network is correct.

In this paper, we therefore perform a validation study of methods to infer gene regulatory relations from expression and sequence information. Using the four methods described above, we predict sigma (transcription) factors in *Bacillus subtilis* from the combined gene expression data of ten time course experiments and 99 gene disruptant experiments for genes whose sigma factor is known experimentally. Sigma factors are transcription factors that bind to the RNA polymerase to enable it to find the appropriate DNA binding sequence upstream of the transcription start site. Without a sigma factor, the RNA polymerase would bind to random sites on the DNA. Here, we consider the sigma factors $\sigma^{\mathrm{D}}$, $\sigma^{\mathrm{E}}$, $\sigma^{\mathrm{G}}$, $\sigma^{\mathrm{H}}$, $\sigma^{\mathrm{K}}$, $\sigma^{\mathrm{L}}$, $\sigma^{\mathrm{W}}$, and $\sigma^{\mathrm{X}}$, which perform particular biological functions in the cell. We do not include the general sigma factors $\sigma^{\mathrm{A}}$ and $\sigma^{\mathrm{B}}$, as well as several minor sigma factors with few known regulated genes.

This particular biological validation study is appropriate for four reasons. First, a sigma factor is needed for transcription for almost all genes in *Bacillus subtilis*. Accordingly, sigma factors tend to regulate a fairly large number of genes, many of which are known for the *Bacillus subtilis* genome, such that a meaningful leave-one-out analysis becomes feasible. Secondly, prokaryotes have simpler mechanisms of gene regulation than eukaryotes. As the biological validity of gene regulatory network inference is not well established, it is appropriate to first analyze a simpler prokaryotic system instead of a eukaryotic system. Third, a large amount of gene expression data is available for *Bacillus subtilis*. Lastly, as in prokaryotes genes belonging to the same operon are transcribed into a single mRNA molecule, we can average the gene expression ratios over each operon to reduce the adverse effects of noise in the measurements.

In this work, we found that Bayesian network and dynamic models fail to accurately predict gene regulation by sigma factors, while coregulation is about 76% accurate in a leave-one-out analysis. Although sequence motif information by itself yields a prediction accuracy of 73%, combining gene expression data and sequence motif information by adding their likelihood scores, as proposed by Segal *et al.* (2003), barely improved the prediction accuracy. However, using a logistic regression model to combine the two likelihood scores yielded a better balance between gene expression data and sequence motif information and resulted in a prediction accuracy of 85% in a leave-one-out analysis.

Using the score functions derived by logistic regression, we searched the complete *Bacillus subtilis* genome for additional genes that are regulated by the sigma factors under consideration. We calculate the logistic regression scores for genes known not to be regulated by a given sigma factor to assess the statistical significance of the newly predicted regulatory relations. By providing both the tentative sigma factor as well as the predicted binding site of the RNA polymerase-sigma factor complex, we enable a straightforward experimental verification of our prediction results.

Lastly, we note that a single paper in the biological literature typically describes one experiment in which only one gene regulatory relation is demonstrated. The gene regulatory relations newly predicted in this paper therefore demonstrate the power of genome-wide data to reveal the gene regulatory network.

## METHODS

### Fold-change analysis

In a fold-change analysis, we calculate by which factor the expression of a particular gene changes following the disruption of a transcription factor. Here, we consider the change in the gene expression if one of the sigma factors is disrupted. The sigma factor whose disruption leads to the largest decrease in the expression of the regulated gene is predicted to drive the transcription of that gene.

### Dynamic models

Dynamic models describe time-course data only, taking the time information explicitly into account. Several dynamic models of gene regulatory networks inferred from time-course gene expression data have been suggested previously. Murphy & Mian (1999) showed that most of the existing discrete time models can be considered as special cases of the general class of dynamic Bayesian networks. Here, we derive a dynamic model from a set of stochastic differential equations (Chen *et al.*, 1999; De Hoon *et al.*, 2003), as they allow a convenient treatment of gene expression measurements made at unequal time intervals.

In a stochastic differential equation model, the rate of change of the gene expression levels $\frac{\mathrm{d}}{\mathrm{d}t}\underline{x}(t)$ at time $t$ is a function of the expression levels $\underline{x}(t)$ at that time point plus a noise term:

$$\frac{\mathrm{d}}{\mathrm{d}t}\underline{x}(t) = \underline{g}(\underline{x}(t)) + \underline{\underline{\dot{\sigma}}} \cdot \underline{\varepsilon}(t), \qquad (1)$$

where the function $\underline{g}$ effectively describes the gene regulatory network, $\underline{\varepsilon}(t)$ is a random process with unit variance, and $\underline{\underline{\dot{\sigma}}} = \mathrm{diag}(\dot{\sigma}_1, \ldots, \dot{\sigma}_m)$ is a diagonal matrix with units of $[\mathrm{time}]^{-1}$. The differential equation can be approximated by a difference equation:

$$\frac{\underline{x}_{i+1} - \underline{x}_i}{t_{i+1} - t_i} = \underline{g}(\underline{x}_i) + \underline{\underline{\dot{\sigma}}} \cdot \underline{\varepsilon}_i, \qquad (2)$$

For measurements taken at equal time intervals ($\Delta t = t_{i+1} - t_i$ independent of $i$), this reduces to a dynamic Bayesian network (Kim *et al.*, 2003). Ong *et al.* (2002) consider a similar model, in which the gene expression data are discretized to binary values, and the gene interactions are described by conditional probability tables. The deterministic models proposed previously (Liang *et al.*, 1998; Akutsu *et al.*, 1999, 2000) are based on Eq. (2) without the error term, after discretizing the expression data. The model proposed by Van Someren *et al.* (2000) ($\underline{x}_{i+1} = \underline{\underline{M}} \cdot \underline{x}_i$ where $\underline{\underline{M}}$ is a square matrix) can be regarded as a special case of Eq. (2), after dropping the noise term and assuming equal time intervals and linear interactions.

In our implementation of a dynamic model, we chose for a linear model with continuous variables. In the validation study described below, a nonlinear dynamic model (Kim *et al.*, 2003) yielded less accurate predictions of gene regulation by sigma factors. This may be due to the larger number of parameters that need to be estimated in a nonlinear model, leading to a less accurate parameter estimation than in a linear model. In this paper, we therefore restrict ourselves to linear models.

### Bayesian networks

We denote the joint probability distribution of the gene expression levels $x_{j,i}$, $j \in \{1, \ldots, m\}$ of $m$ genes measured in experiment $i$ as $P(x_{1,i}, x_{2,i}, \ldots, x_{m,i})$. In the Bayesian network, we assume that this joint probability distribution can be decomposed as

$$P\left(x_{1,i}, x_{2,i}, \ldots, x_{m,i}\right) = \prod_{j=1}^{m} P_j\left(x_{j,i} \middle| \left\{x_{j',i}; j' \in \mathrm{Pa}\left(j\right)\right\}\right), \quad (3)$$

where $\mathrm{Pa}\left(j\right)$ represents the set of parent genes (regulators) of gene $j$. This decomposition can then be represented as a directed acyclic graph.

To apply this formula in practice, we need to choose an appropriate mathematical form for the gene regulations encoded by the right hand side of this equation. Friedman *et al.* (2000) proposes to either discretize the gene expression data and represent their dependencies as a truth table, or use continuous variables whose dependencies are described by linear relations. To avoid the information loss associated with discretizing gene expression data, we chose the latter option. The Bayesian network model then essentially looks for linear correlations between parent genes and child genes. The Bayesian network can be applied to expression data from both gene disruptant and time course experiments, though in the latter case no use is made of the time information.

### Inference based on coregulation

The three inference methods described above consider the parent gene directly to discover gene regulatory relations. We may also be able to find gene regulatory relations by comparing the gene expression profiles of different child genes to each other. This approach is usually applied in an unsupervised setting, in which gene expression data are clustered based on the similarity in their gene expression profile. If, for a given transcription factor, a large number of regulated genes are already known, we can also predict gene regulatory relations by comparing the gene expression profiles of genes in the same regulon to the gene expression profile of a new gene. We can then infer gene regulatory relations in a supervised setting by making use of known regulatory relations.

Segal *et al.* (2003) describes the gene expression measurements of coregulated genes by a normal distribution, assuming that measurements in the $n$ different experiments or time points are statistically independent:

$$p^{(s)}\left(x_{j,1}, x_{j,2}, \ldots, x_{j,n}\right) = \prod_{i=1}^{n} p_i^{(s)}\left(x_{j,i}\right). \quad (4)$$

Here, $x_{j,i}$ is the expression log-ratio measured in experiment $i$ of gene $j$ regulated by sigma factor $s$, and $p_i^{(s)}\left(x_{j,i}\right)$ is a normal distribution:

$$p_i^{(s)}\left(x_{j,i}\right) = \frac{1}{\sigma_i^{(s)}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_{j,i}-\mu_i^{(s)}}{\sigma_i^{(s)}}\right)^2\right]. \quad (5)$$

For the regulon of each sigma factor $s$, we then calculate the mean $\mu_i^{(s)}$ and standard deviation $\sigma_i^{(s)}$ in each experiment $i$, and calculate the log-likelihood of a new gene, given its expression measurements $y_i$, to belong to the same regulon as

$$L_{\mathrm{expr}}^{(s)}\left(y_1, y_2, \ldots, y_n\right) = -\frac{n}{2}\ln\left(2\pi\right) - \sum_{i=1}^{n}\ln\sigma_i^{(s)} - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\mu_i^{(s)}}{\sigma_i^{(s)}}\right)^2. \quad (6)$$

This likelihood score is calculated for the regulon of each sigma factor $s$ to determine which regulon agrees best in terms of gene expression with the gene expression profile of the new gene.

In practice, we found that due to the reduced effect of outliers, estimating the standard deviation $\sigma$ from the combined experiments via

$$\sigma^{(s)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\sigma_i^{(s)}\right)^2}, \quad (7)$$

yielded a more accurate prediction of regulation by sigma factors. We therefore applied Eq. (6) with $\sigma_i^{(s)}$ replaced by $\sigma^{(s)}$ in all cases.

### Motif search

In addition to the gene expression data, we may make use of the sequence motif information of the RNA polymerase-sigma factor DNA binding site. The motifs of these binding sites for sigma factors consists of two parts, one located around 35 base pairs and another around 10 base pairs upstream of the transcription start site. The distance between the transcription start site and the translation start site varies, but is generally not more than about 300 base pairs. The gap between the -35 and the -10 binding motifs can differ for different genes in the same regulon, but not by more than one or two base pairs or so.

| Sigma factor | Binding motif |
|---|---|
| $\sigma^{\mathrm{D}}$ | TAAA (13-15) GCCGATATAA |
| $\sigma^{\mathrm{E}}$ | GCATATTT (12-14) CATACAAT |
| $\sigma^{\mathrm{G}}$ | GCATA (17-18) CATACTA |
| $\sigma^{\mathrm{H}}$ | GAAGGAATT (14-15) GAAT |
| $\sigma^{\mathrm{K}}$ | AC (17-19) CATATGAT |
| $\sigma^{\mathrm{L}}$ | TGGCA (5) CTTGCAT |
| $\sigma^{\mathrm{W}}$ | TGAAACCTT (13-14) CGTATA |
| $\sigma^{\mathrm{X}}$ | TGAAAC (16-17) CGTCTA |

**Table 1.** The consensus sequence of the DNA binding motifs for the RNA polymerase-sigma factor binding site for the eight sigma factors under consideration. The left motif is located around 35 base pairs in front of the transcription start site (except for $\sigma^{\mathrm{L}}$), while the right motif is located at about -10 base pairs.

Table 1 shows the consensus motifs for the sigma factors under consideration here, as determined using Bioprospector (Liu *et al.*, 2001) from the DBTBS database of transcriptional binding sequences in *Bacillus subtilis* (Makita *et al.*, 2004). Whereas some sigma factors, such as $\sigma^{\mathrm{L}}$, can be distinguished easily from other sigma factors by virtue of its distinct sequence motif, other sigma factors such as $\sigma^{\mathtt{D}}$ and $\sigma^{\mathtt{E}}$ have similar motifs, which may not be easily distinguished based on motif information alone.

The motif sequences can be described statistically by a position specific score matrix $M_{k,p}^{(s)}$ (Durbin *et al.*, 1998) for sigma factor $s$, which lists the log-odds score of finding a nucleotide $p$ at position $k$ in the binding sequence motif of sigma factor $s$. The log-likelihood, relative to the background sequence probabilities, for a sequence $S[k]$ is then

$$L_{\mathrm{motif}}^{(s)}(S) = \sum_{k=1}^{K} M_{k,S[k]}^{(s)}, \qquad (8)$$

where $K$ is the length of the motif. For the sequence motifs for RNA polymerase-sigma factor binding sites, we added the score of the -35 and the -10 motifs, and allowed the gap to vary according to the currently known binding sites.

The position specific score matrix was calculated from the known binding motifs of the genes in the regulon of each sigma factor, as listed in the DBTBS database. For the matrix calculation based on $N$ known binding sites, we added $\sqrt{N}$ pseudocounts, using a background probability of 0.3185 for A and T, and 0.1815 for C and G.

### Combining gene expression and motif information

Segal *et al.* (2003) proposed to add the log-likelihood scores based on the gene expression data and the motif information into a single log-likelihood score:

$$L^{(s)} = L_{\mathrm{expr}}^{(s)}(y_1, \ldots, y_n) + L_{\mathrm{motif}}^{(s)}(S). \qquad (9)$$

Here, $L_{\mathrm{motif}}^{(s)}(S)$ is the log-likelihood score for the highest-scoring sequence motif $S$ in the 300 base pair region upstream of the translation start site. By combining the two information sources, we expect to be able to gain a higher prediction accuracy. For sigma factors such as $\sigma^{\mathrm{L}}$, having a distinctive sequence motif, we expect the second term to be dominant, while the gene expression score may help us to distinguish $\sigma^{\mathrm{D}}$, $\sigma^{\mathrm{E}}$, which have similar sequence motifs. We will revisit this equation below, where we show that Eq. (9) does not achieve the optimal balance between gene expression data and sequence information, and may even essentially ignore the latter.

## ASSESSMENT OF BIOLOGICAL VALIDITY

Bayesian network models have been predicted previously from measured gene expression data of *Saccharomyces cerevisiae* (Imoto *et al.*, 2002a; Kim *et al.*, 2003; Tamada *et al.*, 2003; Friedman *et al.*, 2000) and *E. coli* (Ong *et al.*, 2002). The validity of those network predictions was assessed heuristically by showing an example of a gene regulatory relation that was found correctly by the model. However, for a useful prediction of gene regulatory relations, we need to know how many of the hundreds or thousands of gene regulatory relations in such an inferred network are correct. While a considerable effort has been aimed at investigating the algorithmic aspects of regulatory network inference, the biological validity of the inferred networks has not yet been clearly demonstrated.

Here, we consider the sigma factors $\sigma^{\mathrm{D}}$, $\sigma^{\mathrm{E}}$, $\sigma^{\mathrm{G}}$, $\sigma^{\mathrm{H}}$, $\sigma^{\mathrm{K}}$, $\sigma^{\mathrm{L}}$, $\sigma^{\mathrm{W}}$, $\sigma^{\mathrm{X}}$ in *Bacillus subtilis*. A large number of genes have been shown experimentally to be regulated by each of these sigma factors, as listed in the DBTBS database (Makita *et al.*, 2004). For each gene that is currently known to be regulated exclusively by one sigma factor (and possibly by other, non-sigma transcription factors), we calculate a Bayesian network, a dynamic model, and a coregulation-based model from the combined gene expression data of ten time-course experiments (Table 2). Sequence motif information is ignored for now.

The gene expression levels were measured twice at each time point. We calculated the average background noise level for the red (cy5) and green (cy3) channel separately for each data set. Gene expression measurements where the fluorescence level is less than the average background level in either channel are removed from the data set, as they will be dominated by noise. Global normalization is then applied by dividing the measured fluorescence levels of the remaining genes by their sum in each channel. We note that in a previous prediction of the operon structure of *Bacillus subtilis* using these expression data, we found a 77.3% accuracy level (De Hoon *et al.*, 2004), which is a typical accuracy level for operon prediction.

Table 3 shows the frequency that each sigma factor was estimated correctly by each network inference method. The dynamic model yielded 44 correct predictions out of 189, an accuracy of 23%. While it is statistically significant ($p = 2.6 \times 10^{-6}$) to predict the sigma factor correctly for 44 out of 189 genes, given the low rate of accurate predictions the dynamic model is unlikely to be a good predictor of currently unknown gene regulatory relations. Bayesian networks perform somewhat better with a prediction accuracy of 25%. This accuracy level could only be attained when the Bayesian network model was applied to normalized log-ratios; a Bayesian network learned from gene expression ratios directly yielded a much lower prediction accuracy. The coregulation-based prediction yields the highest prediction accuracy at 52% in a leave-one-out analysis, in which for the prediction of a gene $j$ the regulon statistics $\mu_i^{(s)}, \sigma_i^{(s)}$ are recalculated after removing gene $j$ from its regulon.

**Table 3.** Number of correct sigma factor predictions for the dynamic model, the Bayesian network model, and the coregulation-based model. For these predictions, only the time-course gene expression data were used.

| Sigma factor | Total | Dynamic model | Bayesian network | Coregulation-based model |
|---|---|---|---|---|
| $\sigma^{\mathrm{D}}$ | 16 | 2 | 12 | 12 |
| $\sigma^{\mathrm{E}}$ | 51 | 17 | 3 | 21 |
| $\sigma^{\mathrm{G}}$ | 25 | 1 | 0 | 7 |
| $\sigma^{\mathrm{H}}$ | 40 | 5 | 5 | 27 |
| $\sigma^{\mathrm{K}}$ | 23 | 10 | 9 | 13 |
| $\sigma^{\mathrm{L}}$ | 6 | 0 | 4 | 4 |
| $\sigma^{\mathrm{X}}$ | 4 | 0 | 2 | 1 |
| $\sigma^{\mathrm{W}}$ | 24 | 9 | 12 | 14 |
| Total | 189 | 44 | 47 | 99 |
| Percentage | | 23% | 25% | 52% |
| $p$-value | | $3.1 \times 10^{-5}$ | $2.6 \times 10^{-6}$ | $1.0 \times 10^{-39}$ |

**Table 4.** Disrupted gene in each experiment. The genes *degU*, *sigF*, *sigW*, and *veg* were each disrupted in two experiments, as indicated here.

| | | | | | |
|---|---|---|---|---|---|
| *abh* | *cspB* | *iolR* | *rocR* | *sigZ* | *yesS* |
| *abrB* | *ctsR* | *ycsO* | *sacT* | *sinR* | *yhjM* |
| *acoR* | *ydbG* | *lacR* | *senS* | *soj* | *yotL* |
| *ahrC* | *degU* (2×) | *levR* | *sigB* | *splA* | *yqfV* |
| *alsR* | *deoR* | *lexA* | *sigD* | *spo0A* | *ytzE* |
| *ansR* | *yjmH* | *lmrA* | *sigE* | *spo0J* | *yufL* |
| *araR* | *yqkL* | *lrpA* | *sigF* (2×) | *spoIIIC* | *yugG* |
| *azlB* | *gerE* | *lrpC* | *sigG* | *spoIIID* | *yurK* |
| *ccpA* | *glcR* | *yqhN* | *sigH* | *spoVT* | *yvkB* |
| *yyaG* | *glcT* | *mtrB* | *ykoZ* | *tenA* | *yvrH* |
| *ykuM* | *glnR* | *paiA* | *sigL* | *tnrA* | *ywaE* |
| *citR* | *gntR* | *paiB* | *yhdM* | *treR* | *yyaA* |
| *citT* | *gutR* | *ygaG* | *sigV* | *veg* (2×) | *yybA* |
| *codY* | *hpr* | *phoP* | *sigW* (2×) | *xylR* | *yybE* |
| *comA* | *hrcA* | *purR* | *sigX* | *ybbH* | *yydK* |
| *comK* | *hutP* | *pyrR* | *sigY* | *ybfA* | |

To improve the prediction accuracy of the Bayesian network and the coregulation-based approach, we augmented our data set with the gene expression measurements of 99 gene disruptant experiments, listed in Table 4. Both methods were then applied to the gene expression data of the combined 174 microarrays. As shown in Table 5, the prediction accuracy increased for both methods upon adding the gene disruptant data. The Bayesian network model yielded an accuracy of 42%, while the coregulation-based model gave the correct sigma factor prediction for 76% of the genes. The fold-change analysis, based only on the expression data from the gene disruptant experiments in which one of the eight sigma factors was disrupted, yielded a prediction accuracy of 54%.

| Experiment | Measurement time points in minutes |
|---|---|
| Cold shock | 0, 5, 10, 30, 60, 120 |
| Competence | 0, 60, 120, 180, 240, 300, 360 |
| Glucose, glutamine added during sporulation | 0, 60, 120, 180, 240, 300 |
| Glucose limitation | 0, 60, 125, 180, 240 |
| Heat shock | 0, 5, 10, 30, 60 |
| Increased aminoacid availability | 0, 30, 60, 120, 210, 300, 420, 540 |
| Phosphate, glucose starvation | 0, 60, 120, 180, 240, 300, 360, 420 |
| Phosphate limitation | 0, 55, 115, 175, 235, 295 |
| Salt stress | 0, 5, 10, 30, 60 |
| Sporulation | 0, 30, 60, 90, 120, 150, 180, 210, 240, 270, 300, 330, 360, 390, 420, 450, 480, 510, 540 |

**Table 2.** The time points at which expression measurements were made for the ten time-course experiments of *Bacillus subtilis* considered in this paper.

**Table 5.** Number of correct sigma factor predictions using sequence motif information and gene expression information, using both the time-course and the gene-disruptant expression data. (*) For the fold-change analysis, only the gene expression data of the gene disruptant experiments were used in which one of the eight sigma factors was disrupted.

| | | Sequence | Expression data | | | Sequence and expression data | |
|---|---|---|---|---|---|---|---|
| Sigma factor | Total | Motif | Fold-change (*) | Bayesian network | Coregulation | Sum of likelihood scores | Logistic regression |
| $\sigma^{D}$ | 16 | 14 | 15 | 13 | 13 | 13 | 13 |
| $\sigma^{E}$ | 51 | 31 | 35 | 17 | 39 | 39 | 43 |
| $\sigma^{G}$ | 25 | 2 | 16 | 1 | 13 | 13 | 19 |
| $\sigma^{H}$ | 40 | 21 | 33 | 5 | 35 | 35 | 35 |
| $\sigma^{K}$ | 23 | 6 | 6 | 21 | 18 | 18 | 19 |
| $\sigma^{L}$ | 6 | 6 | 6 | 2 | 4 | 4 | 6 |
| $\sigma^{X}$ | 4 | 1 | 3 | 1 | 1 | 2 | 2 |
| $\sigma^{W}$ | 24 | 21 | 24 | 20 | 21 | 22 | 24 |
| Total | 189 | 138 | 102 | 80 | 144 | 146 | 161 |
| Percentage | | 73% | 54% | 42% | 76% | 77% | 85% |

### Combining gene expression and motif information, revisited

The prediction accuracy of the coregulation-based model can be improved further by adding the likelihood scores of the gene expression data and the sequence information (Eq. (9)), as proposed by Segal *et al.* (2003). From Table 5, we see that the sequence information alone gives a prediction accuracy of about 73%, just slightly lower than the combined gene expression data. This is in agreement with the previous result that sequence information by itself can give a rather good prediction of regulation by sigma factors (Yada *et al.*, 1997). We would therefore expect that the combination of gene expression data and sequence motif information would lead to an improved prediction accuracy. However, as Table 5 shows, the combined score achieved a prediction accuracy of 77%, just barely larger than the gene expression data by themselves. It is particularly surprising that the prediction accuracy does not improve for $\sigma^{L}$, whose sequence motif is easily distinguishable from the sequence motifs of the other sigma factors.

The failure to effectively make use of sequence motif information is caused by the assumptions underlying Eq. (9). Particularly, Eq. (4) assumes that the gene expression data from different experiments are statistically independent. This may be a valid assumption if the perceived randomness in the expression data is caused by measurement errors only. However, in reality the variability in the expression measurements, represented by $\sigma_i^{(s)}$, includes both measurement errors as well as biological knowledge that is either unknown or ignored, such as additional transcription factors regulating a gene. In Eq. (4), each additional microarray experiment is then incorrectly assumed to contribute an equal amount of new information as the previous experiments. If the number of microarrays is large, as in our case, the likelihood score of the expression data will overwhelm the sequence motif score, which is then effectively ignored.

A common statistical technique to correct this situation is logistic regression (Hastie *et al.*, 2001). In a logistic regression model, we treat the gene expression score and the sequence motif score as two random variables. The probability of belonging to the regulon of sigma factor $s$ is given by the logistic function:

$$\mathrm{Pr}\ (\text{gene belongs to regulon } s | L_{\mathrm{expr}}, L_{\mathrm{motif}}) = \frac{\exp\left(w_0^{(s)} + w_{\mathrm{expr}}^{(s)} L_{\mathrm{expr}}^{(s)} + w_{\mathrm{motif}}^{(s)} L_{\mathrm{motif}}^{(s)}\right)}{\sum_{s'} \exp\left(w_0^{(s')} + w_{\mathrm{expr}}^{(s')} L_{\mathrm{expr}}^{(s')} + w_{\mathrm{motif}}^{(s')} L_{\mathrm{motif}}^{(s')}\right)}.$$

Accordingly, the log-likelihood score is again the sum of the gene expression score and the sequence motif score, but now each is preceded by a weight:

$$L^{(s)} = w_0^{(s)} + w_{\mathrm{expr}}^{(s)} \cdot L_{\mathrm{expr}}^{(s)} + w_{\mathrm{motif}}^{(s)} \cdot L_{\mathrm{motif}}^{(s)} - \ln\left[\sum_{s'} \exp\left\{w_0^{(s')} + w_{\mathrm{expr}}^{(s')} \cdot L_{\mathrm{expr}}^{(s')} + w_{\mathrm{motif}}^{(s')} \cdot L_{\mathrm{motif}}^{(s')}\right\}\right]$$

The weights can be estimated by maximizing this likelihood score, given the gene expression score and sequence motif scores for the genes whose sigma factor is known. As the likelihood score is a nonlinear function of the weights, maximizing this score is not straightforward. However, we found that in practice a simple Newton-Raphson method starting from zero weights converges quickly.

The weights were recalculated each time a gene was removed for the leave-one-out analysis of this weighted score function. As shown in Table 5, the logistic regression score yielded an improved prediction accuracy of 85%. As expected, the logistic regression model is able to recover the 100% accuracy rate for the $\sigma^{\mathrm{L}}$ transcription factor. For $\sigma^{\mathrm{E}}$, $\sigma^{\mathrm{H}}$, and $\sigma^{\mathrm{K}}$, for which the gene expression data gave a more accurate prediction than the sequence motif information, the logistic regression score maintains the prediction accuracy of the gene expression data for $\sigma^{\mathrm{H}}$, and yields a prediction accuracy that surpasses those of the motif information and the gene expression data separately for $\sigma^{\mathrm{E}}$ and $\sigma^{\mathrm{K}}$. As shown in Table 5, adding the likelihood scores directly did not improve the prediction accuracy for $\sigma^{\mathrm{E}}$, $\sigma^{\mathrm{H}}$, and $\sigma^{\mathrm{K}}$.

## GENOME-WIDE SEARCH FOR GENES REGULATED BY SIGMA FACTORS

We calculated the score function based on logistic regression for all operons in the *Bacillus subtilis* genome in order to find currently unknown gene regulations by sigma factors. As for the most part the operon structure of *Bacillus subtilis* has not been determined experimentally, we use the computationally predicted operon structure instead (De Hoon *et al.*, 2004), leading to a total of 2214 operons.

The logistic regression score was calculated for each sigma factor, both for operons that are known to belong to the corresponding regulon, and for operons known to be regulated by other sigma factors. This allows us to calculate the $p$-value of the logistic regression score for newly predicted gene regulations by a given sigma factor, under the null hypothesis that a gene is not regulated by that sigma factor.

Table 6 shows the currently unknown predicted gene regulations by sigma factors, for which the predicted score was statistically significant at a significance level of $\alpha = 5 \times 10^{-4}$. These genes are characterized by both a high similarity in gene expression with other genes regulated by the predicted sigma factor, and a binding sequence motif that is highly consistent with the consensus sequence. With the putative DNA binding sequences available, an experimental validation of these predictions should be straightforward.

Of particular interest is the *yusZ-mrgA* operon and the *yvbX-yvbY-yvfW-yvfV-yvfU-yvfT* operon, which are both predicted to be regulated by $\sigma^{\mathrm{L}}$. So far, only six operons are known experimentally to be regulated by $\sigma^{\mathrm{L}}$. These six operons

| Operon | Sigma factor | Motif | Approximate distance between transcription and translation start sites |
|---|---|---|---|
| *ybdO* | $\sigma^{\mathrm{D}}$ | **TAA**T — 15 bp — **GCCGATA**A**AA** | 25 |
| *deoR-yxxB-yxeR* | $\sigma^{\mathrm{D}}$ | **TAA**C — 13 bp — **GCCGATATAA** | 85 |
| *yqjV-yqjU* | $\sigma^{\mathrm{D}}$ | **T**CAT — 13 bp — **GCCGATAT**G**A** | 250 |
| *yusZ-mrgA* | $\sigma^{\mathrm{L}}$ | **TGGC**C — 5 bp — **CTTGCA**G | 130 |
| *yvbX-yvbY-yvfW-yvfV-yvfU-yvfT* | $\sigma^{\mathrm{L}}$ | **TGGC**C — 5 bp — **CTT**C**C**G**T** | 265 |
| *ypuA* | $\sigma^{\mathrm{W}}$ | **TGAAACCTG**C — 14 bp — **CGTCTA** | 80 |

**Table 6.** Newly predicted gene regulations by sigma factors in *Bacillus subtilis*. These predictions are statistically significant to a level of $\alpha = 5 \times 10^{-4}$. In the motifs, bold characters are consistent with the consensus motif (compare to Table 1).

have identical binding motifs, except for *rocG*, whose binding sequence **TGG**T**A** — 5 bp — **CTTGCAT** deviates in one position from the consensus motif. Our newly predicted operon *yusZ-mrgA* with the binding sequence motif **TGGC**C — 5 bp — **CTTGCA**G deviates in two places, while the binding motif of *yvbX-yvbY-yvfW-yvfV-yvfU-yvfT*, **TGGC**C — 5 bp — **CTT**C**C**G**T**, deviates in three positions. The strong similarity in gene expression to other $\sigma^{\mathrm{L}}$ regulated genes, together with the motif similarity, leads to a highly significant prediction. A simple experimental verification of these predictions is possible with the predicted binding sites of the $\sigma^{\mathrm{L}}$-RNA polymerase binding sites listed in Table 6.

## DISCUSSION

To our knowledge, this is the first thorough assessment of the biological validity of gene regulatory relations inferred from genome-wide data. We found that coregulation is a considerably better predictor of gene regulation by sigma factors in *Bacillus subtilis* than Bayesian networks, dynamic models, or a fold-change analysis. A Bayesian network performs slightly better than a dynamic model, particularly because it allows the use of both time-course gene expression data and gene disruptant information. A fold-change analysis, though based on a much smaller amount of gene expression data, performs better than Bayesian networks and dynamic models. However, a fold-change analysis is possible only if a disruption experiment for the transcription factor under consideration is available, while Bayesian networks, dynamic models, and coregulation do not have this requirement.

The superior performance of sigma factor prediction from coregulation is likely due to the larger amount of expression data on which it depends. For example, the regulon of $\sigma^{\mathrm{E}}$ in our study contains 51 genes, whereas Bayesian networks and dynamic models make use of the expression data of the transcription factor only. Here, we were able to make use of coregulation in a supervised fashion because of the large number of regulated genes known for each sigma factor. When the aim is to find new transcription factors, it will be necessary to consider the gene expression data of the parent gene directly, either by a Bayesian network, a dynamic model, or a fold-change analysis. Currently, such models predict gene regulations based on the parent-child relation only. Given our prediction accuracies, it may be advisable to include similarity to coregulated genes explicitly in these models.

The prediction accuracies can be improved further by including DNA sequence motif information as an additional predictor in the model, as recently proposed by Segal *et al.* (2003) and Tamada *et al.* (2003). It is important to balance the gene expression and the sequence motif information carefully to optimize the predictive power of the joint score. As we have shown, simply adding the log-likelihood scores for each predictor (Segal *et al.*, 2003) effectively ignores sequence information if the number of microarrays is large. Instead, we estimate the relative predictive power of gene expression and sequence motif information from the data themselves using a logistic regression model, leading to an effective use of both information sources and an improved prediction accuracy.

We note that the logistic regression model typically increased the relative importance of the sequence information by about a factor of ten compared to directly adding the log-likelihood scores.

We then performed a genome-wide search using the score function derived from the logistic regression model to find additional genes that are regulated by each sigma factor. A very high score was found for several genes, for which we can be confident that our predictions are correct. For genes with lower scores, it becomes progressively more difficult to decide if the prediction is correct or if the score is based on chance. However, since our method identifies the location of the binding site of the sigma factor-RNA polymerase complex as part of the sigma factor prediction, a simple experimental verification of the predictions is possible.

## 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. In *Proc. Pac. Symp. on Biocomputing*, volume 5. pp. 17–28.

Akutsu, T., Miyano, S. & Kuhara, S. (2000). Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, **16**, 727–734.

Chen, T., He, H. L. & Church, G. M. (1999). Modeling gene expression with differential equations. In *Proc. Pac. Symp. on Biocomputing*, volume 4. pp. 29–40.

De Hoon, M., Imoto, S., Kobayashi, K., Ogasawara, N. & Miyano, S. (2003). Inferring gene regulatory networks from time-ordered gene expression data of *Bacillus Subtilis* using differential equations. In *Proc. Pac. Symp. on Biocomputing*, volume 8. pp. 17–28.

De Hoon, M., Imoto, S., Kobayashi, K., Ogasawara, N. & Miyano, S. (2004). Predicting the operon structure of *Bacillus Subtilis* using operon length, intergene distance, and gene expression information. In *Proc. Pac. Symp. on Biocomputing*, volume 9. pp. 276–287.

Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998). Biological sequence analysis. Cambridge University Press, Cambridge, UK.

Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**, 601–620.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York.

Imoto, S., Goto, T. & Miyano, S. (2002a). Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. In *Proc. Pac. Symp. on Biocomputing*, volume 7. pp. 175–186.

Imoto, S., Kim, S., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S. & Miyano, S. (2002b). Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. In *IEEE Computer Society Bioinformatics Conference (CSB2002)*. pp. 219–227.

Kim, S., Imoto, S. & Miyano, S. (2003). Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. In *International Workshop on Computational Methods in Systems Biology (CMSB2003), Springer Verlag Lecture Notes in Computer Science*, volume 2602. pp. 104–113.

Liang, S., Fuhrman, S. & Somogyi, R. (1998). REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Proc. Pac. Symp. on Biocomputing*, **3**, 18–29.

Liu, X., Brutlag, D. & Liu, J. (2001). Bioprospector: Discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. In *Proc. Pac. Symp. on Biocomputing*, volume 6. pp. 127–38.

Makita, Y., Nakao, M., Ogasawara, N. & Nakai, K. (2004). DBTBS: Database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Research*, **32**, D75–D77. http://dbtbs.hgc.jp.

Murphy, K. & Mian, S. (1999). Modelling gene expression data using dynamic Bayesian networks. Technical report, University of California, Berkeley.

Ong, I. M., Glasner, J. D. & Page, D. (2002). Modelling regulatory pathways in E. coli from time series expression profiles. In *Proceedings of the Tenth International Conference on Intelligent Systems for Molecular Biology (ISMB 2002), Bioinformatics Supplement 1*. pp. 241–248.

Pe'er, D., Regev, A., Elidan, G. & Friedman, N. (2001). Inferring subnetworks from perturbed expression profiles. In *Proceedings of the Ninth International Conference on Intelligent Systems for Molecular Biology (ISMB 2001), Bioinformatics Supplement 1*. pp. 215–224.

Segal, E., Yelensky, R. & Koller, D. (2003). Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. In *Proceedings of the Ninth International Conference on Intelligent Systems for Molecular Biology (ISMB 2003), Bioinformatics Supplement 1*. pp. i273–i282.

Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S. & Miyano, S. (2003). Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. In *Proceedings of the Second European Conference on Computational Biology*. pp. ii227–ii236.

Van Someren, E. P., Wessels, L. F. A. & Reinders, M. J. T. (2000). Linear modeling of genetic networks from experimental data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, volume 8. pp. 355–366.

Yada, T., Totoki, Y., Ishii, T. & Nakai, K. (1997). Functional prediction of *Bacillus subtilis* genes from their regulatory sequences. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*. pp. 354–357.