



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

M. J. L. de Hoon^{1,*}, Y. Makita¹, S. Imoto¹, K. Kobayashi²,
N. Ogasawara², K. Nakai¹ and S. Miyano¹

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

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Motivation: Sigma factors regulate the expression of genes in *Bacillus subtilis* at the transcriptional level. We assess the accuracy of a fold-change analysis, Bayesian networks, dynamic models and supervised learning based on coregulation in predicting gene regulation by sigma factors from gene expression data. To improve the prediction accuracy, we combine sequence information with expression data by adding their log-likelihood scores and by using a logistic regression model. We use the resulting score function to discover currently unknown gene regulations by sigma factors.

Results: The coregulation-based supervised learning method gave the most accurate prediction of sigma factors from expression data. We found that the logistic regression model effectively combines expression data with sequence information. In a genome-wide search, highly significant logistic regression scores were found for several genes whose transcriptional regulation is currently unknown. We provide the corresponding RNA polymerase binding sites to enable a straightforward experimental verification of these predictions.

Contact: mdehoon@ims.u-tokyo.ac.jp

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 gene expression levels 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 or time points.

Alternatively, Bayesian networks inferred from cDNA microarray data have been proposed as a model of gene regulation (Friedman *et al.*, 2000; Imoto *et al.*, 2002, 2003). A Bayesian network shows how the expression level of each gene depends conditionally on a small set of parent genes. 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. 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 by comparing the expression profile of a new gene to the expression profiles of the genes known to be regulated by the 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 a model based on coregulation in which sequence information and gene expression data are combined in a single Bayesian score function;

*To whom correspondence should be addressed.

Tamada *et al.* (2003) proposed a similar method in the framework of Bayesian networks. Within the context of supervised learning to predict transcription factors, this approach reduces to adding the log-likelihood scores for the sequence motif and the gene expression data. In this work, we found that in practice this may lead to an overestimation of the predictive power of gene-expression data to the degree that the sequence motif information is effectively ignored. To find a better balance between sequence information and gene expression data, we propose a logistic regression model to combine the two predictors.

Whereas the algorithmic aspects of methods to infer gene regulatory relations have been well studied in the past, it is still unknown if they yield biologically correct results. Previous biological support of these methods has been limited to finding 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 know beforehand which of the large number of predicted gene regulatory relations in an inferred network are 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 10 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. Here, we consider the sigma factors σ^D , σ^E , σ^G , σ^H , σ^K , σ^L , σ^W and σ^X which perform particular biological functions in the cell. We do not include the general sigma factors σ^A and σ^B , as well as several minor sigma factors with few known regulated genes.

This particular biological validation study is appropriate for three reasons. First, a sigma factor is needed for transcription for almost all genes in *B.subtilis*. Accordingly, sigma factors tend to regulate a fairly large number of genes, many of which are known for the *B.subtilis* genome, such that a meaningful leave-one-out analysis becomes feasible. Second, 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 *B.subtilis*.

In this work, we found that Bayesian networks and dynamic models fail to accurately predict gene regulation by sigma factors, while coregulation was $\sim 76\%$ accurate in a leave-one-out analysis. Although sequence motif information by itself yielded a prediction accuracy of 73%, combining gene expression data and sequence motif information by adding their

likelihood scores barely improved the prediction accuracy. On the other hand, the logistic regression model, by combining the two likelihood scores more effectively, yielded a prediction accuracy of 85% in a leave-one-out analysis.

Using the score functions derived by logistic regression, we searched the complete *B.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.

METHODS

Fold-change analysis

In a fold-change analysis, we calculate by what 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 gene regulatory networks inferred from time-course gene expression data, taking the time information explicitly into account. Several dynamic models have been suggested previously. Murphy and 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 $d\bar{x}(t)/dt$ at time t is a function of the expression levels $\bar{x}(t)$ at that time point plus a noise term:

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

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

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

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) are equivalent to Equation (2) without the error term, after discretizing the expression data. The model proposed by Van Someren *et al.* (2000) ($x_{i+1} = \underline{R} \cdot x_i$ where \underline{R} is a square matrix) can be regarded as a special case of Equation (2), after dropping the noise term and assuming equal time intervals and linear interactions.

In this paper, we also use a linear model [$g(x) = \underline{\Delta} \cdot x$, where $\underline{\Delta}$ is a square matrix], but we allow unequal time intervals between measurements. In the validation study described below, a non-linear 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 non-linear model, leading to a less accurate parameter estimation.

Bayesian networks

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

$$P(x_{1,i}, \dots, x_{m,i}) = \prod_{j=1}^m P_j(x_{j,i} | \{x_{j',i}; j' \in \text{Pa}(j)\}), \quad (3)$$

where $\text{Pa}(j)$ 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 Equation (3). Friedman *et al.* (2000) proposes to either discretize the gene expression data and represent their dependencies as a probabilistic 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 correlations in the expression profile between parent and child genes. A 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 profiles. If, for a given transcription factor, a large number of regulated genes are already known, we can 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. Gene regulatory relations can then be inferred in a supervised manner 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:

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

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

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

For the regulon of each sigma factor s , we then calculate the mean $\mu_i^{(s)}$ and SD $\sigma_i^{(s)}$ 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

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

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 SD from the combined experiments via

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

yielded a more accurate prediction of regulation by sigma factors. We therefore applied Equation (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 consist of two parts, one located ~ 35 bp and another ~ 10 bp upstream of the transcription start site.

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

Sigma factor	Binding motif		
σ^D	TAAA	(13–15 bp)	GCCGATATAA
σ^E	GCATATTT	(12–14 bp)	CATACAAT
σ^G	GCATA	(17–18 bp)	CATACTA
σ^H	GAAGGAATT	(14–15 bp)	GAAT
σ^K	AC	(17–19 bp)	CATATGAT
σ^L	TGGCA	(5 bp)	CTTGCAAT
σ^W	TGAAACCTT	(13–14 bp)	CGTATA
σ^X	TGAAAC	(16–17 bp)	CGTCTA

The left motif is located ~ 35 bp in front of the transcription start site (except for σ^L), while the right motif is located at about -10 bp.

The distance between the transcription start site and the translation start site varies, but is generally not more than ~ 300 bp. The gap between the -35 and the -10 binding motifs can differ by 1 or 2 bp or so for different genes in the same regulon.

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 *B. subtilis* (Makita *et al.*, 2004). Whereas some sigma factors, such as σ^L , can be distinguished easily from other sigma factors by virtue of its distinctive sequence motif, other sigma factors have similar motifs and are not easily distinguishable 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 . This matrix 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

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

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

For the coregulation-based model, a single log-likelihood score based on both the gene-expression data and the motif information can be found by adding their log-likelihood scores

directly:

$$L^{(s)} = L_{\text{expr}}^{(s)}(y_1, \dots, y_n) + L_{\text{motif}}^{(s)}(S). \quad (9)$$

Here, $L_{\text{motif}}^{(s)}(S)$ is the log-likelihood score for the highest-scoring sequence motif S in the 300-bp region upstream of the translation start site, while the expression score $L_{\text{expr}}^{(s)}(y_1, \dots, y_n)$ is calculated from Equation (6). By combining the two information sources, we aim to achieve a higher prediction accuracy and to distinguish sigma factors such as σ^L , having a distinctive sequence motif, based on the motif score, and sigma factors with similar sequence motifs by their gene expression score.

In practice, we found that Equation (9) fails to effectively make use of sequence motif information, as shown below in the leave-one-out analysis. The failure is caused by the assumption 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. As Equation (4) assumes each additional microarray experiment to contribute an equal amount of new information as the previous experiments, the likelihood score of the expression data will overwhelm the sequence motif score if the number of microarrays is large compared with the motif length ($n \gg K$), as in our case. The sequence motif score 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

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

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:

$$\begin{aligned} L^{(s)} &= w_0^{(s)} + w_{\text{expr}}^{(s)} \cdot L_{\text{expr}}^{(s)} + w_{\text{motif}}^{(s)} \cdot L_{\text{motif}}^{(s)} \\ & - \ln \left[\sum_{s'} \exp \left\{ w_0^{(s')} + w_{\text{expr}}^{(s')} \cdot L_{\text{expr}}^{(s')} + w_{\text{motif}}^{(s')} \cdot L_{\text{motif}}^{(s')} \right\} \right]. \end{aligned} \quad (10)$$

The weights can be estimated by maximizing this likelihood score, given the gene expression score and sequence motif

Table 2. The time points at which expression measurements were made for the 10 time-course experiments of *B. subtilis* considered in this paper

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 amino acid 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

scores for the genes whose sigma factor is known. The maximization may not be straightforward, as the likelihood score is a non-linear function of the weights. However, we found that in practice a simple Newton–Raphson method starting from zero weights converges quickly.

ASSESSMENT OF BIOLOGICAL VALIDITY

We assess the biological validity of currently available methods of inferring gene regulatory relations, described above, by evaluating their ability to predict gene regulation by the sigma factors σ^D , σ^E , σ^G , σ^H , σ^K , σ^L , σ^W and σ^X in *B. subtilis*. A large number of genes have been experimentally shown 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 10 time-course experiments (Table 2), ignoring sequence motif information for now.

In these time-course experiments, the gene expression levels were measured twice at each time point. Similarly, in the gene disruptant experiments described below, the expression levels were measured twice in each condition. The average background noise level was calculated separately for the red (cy5) and green (cy3) channel 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. To reduce the effect of noise, we averaged the gene-expression data over each operon.

Table 3 shows the frequency that each sigma factor was predicted correctly by each network inference method. The

Table 3. Number of correct sigma-factor predictions for the dynamic model, the Bayesian network model and the coregulation-based model

Sigma factor	Total	Dynamic model	Bayesian network	Coregulation-based model
σ^D	16	6	12	12
σ^E	51	23	3	21
σ^G	25	1	0	7
σ^H	40	9	5	27
σ^K	23	8	9	13
σ^L	6	0	4	4
σ^W	24	10	12	14
σ^X	4	1	2	1
Total	189	58	47	99
Percentage	–	31	25	52
<i>p</i> -value	–	3.9×10^{-11}	2.6×10^{-6}	1.0×10^{-39}

For these predictions, only the time-course gene expression data were used.

dynamic model yielded 58 correct predictions out of 189, an accuracy of 31%. While predicting the sigma factor correctly for 58 out of 189 genes is statistically significant ($p = 3.9 \times 10^{-11}$), 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 even more poorly with a prediction accuracy of 25%. This accuracy level is 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.

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 predicted the correct sigma factor 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%. Note that the coregulation-based model is a supervised method, while the other three methods are unsupervised.

The prediction accuracy of the coregulation-based model can be improved further by considering the sequence motif information. From Table 5, we see that the sequence information alone gives a prediction accuracy of $\sim 73\%$, just slightly lower than the combined gene expression data. This is

Table 4. Disrupted gene in each of the gene disruptant experiments

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

The genes *degU*, *sigF*, *sigW* and *veg* were each disrupted in two experiments, as indicated here.

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). However, as Table 5 shows, directly adding the likelihood scores of gene expression data and the sequence information [Equation (9)] results in a prediction accuracy of 77%, just barely larger than the prediction based on the expression data only. In particular, the prediction accuracy does not improve for σ^L , despite its distinctive sequence motif.

The logistic regression model of Equation (10) improves the balance between the gene expression data and the sequence motif information by weighting the two predictors. 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 capable of recovering the 100% accuracy rate for the σ^L transcription factor. The logistic regression score maintains the prediction accuracy of the gene expression data for σ^H , for which the gene expression data gave a more accurate prediction than the sequence motif information. The prediction accuracy of the logistic regression score surpasses those of the motif information and the gene expression data separately for σ^E and σ^K .

GENOME-WIDE SEARCH FOR GENES REGULATED BY SIGMA FACTORS

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

Calculating for each sigma factor the logistic regression scores of operons known to be regulated by one of the other sigma factors allowed us to calculate the *p*-value of the logistic regression score for a newly predicted gene regulation by a given sigma factor, under the null hypothesis that the gene is not regulated by this sigma factor. Table 6 shows the newly predicted gene regulations by sigma factors, for which the prediction score was statistically significant at a significance level of $\alpha = 1 \times 10^{-6}$. These genes are characterized by both a high similarity in gene expression profile with other genes regulated by the 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 are the *yusZ-mrgA* and the *ypkP-dfrA-thyB-ypjQ-ypjP* operons, which are both predicted to be regulated by σ^L . So far, only six operons are known experimentally to be regulated by σ^L . These six operons have identical binding motifs, except for *rocG*, whose binding sequence deviates in one position from the consensus motif. Our newly predicted operons *yusZ-mrgA* and *ypkP-dfrA-thyB-ypjQ-ypjP* each deviate in two positions. An experimental validation may reveal if these two operons are indeed regulated by σ^L , in spite of the larger deviation from the consensus binding motif.

DISCUSSION

We found that coregulation is a considerably better predictor of gene regulation by sigma factors in *B.subtilis* than Bayesian networks, dynamic models or a fold-change analysis. The dynamic model performed better than a Bayesian network when applied to time-course gene expression data only; however, the Bayesian network performed better than a dynamic model when applied to both time-course gene expression data and gene disruptant data. A fold-change analysis, though based on a much smaller amount of gene expression data, achieved a better accuracy 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 σ^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

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

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

^aFor 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.

^bUsing the coregulation-based model [Equation (6)] for the expression data.

Table 6. Newly predicted gene regulations by sigma factors in *B.subtilis*

Operon	Sigma factor	Motif	Approximate distance between transcription and translation start sites
<i>deoR-yxxB-yxeR</i>	σ^D	TAAC —13 bp— GCCGATATAA	90
<i>ybdO</i>	σ^D	TAAT —15 bp— GCCGATAAAA	25
<i>ypuA</i>	σ^W	TGAAACCTG —14 bp— CGTCTA	80
<i>ypkP-dfrA-thyB-ypjQ-ypjP</i>	σ^L	TAGTA —5 bp— CTTGCAT	55
<i>yqjV-yqjU</i>	σ^D	TCAT —13 bp— GCCGATATGA	250
<i>yusZ-mrgA</i>	σ^L	TGGCC —5 bp— CTTGCAG	130

These predictions are statistically significant to a level of $\alpha = 1 \times 10^{-5}$. In the motifs, boldface characters are consistent with the consensus motif (compare to Table 1).

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, directly adding the log-likelihood scores for each predictor may effectively ignore the 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, as shown by the improved prediction accuracy. We note that the logistic regression model typically increased the relative importance of the sequence information by about a factor of 10.

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, as our method identifies the location of the DNA-binding site of the sigma factor–RNA polymerase complex as part of the sigma-factor prediction, a straightforward experimental verification of the predictions is possible.

REFERENCES

- Akutsu, T., Miyano, S. and Kuhara, S. (1999) Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac. Symp. Biocomput.*, **4**, 17–28.
- Chen, T., He, H.L. and Church, G.M. (1999) Modeling gene expression with differential equations. *Pac. Symp. Biocomput.*, **4**, 29–40.
- De Hoon, M.J.L., Imoto, S., Kobayashi, K., Ogasawara, N. and Miyano, S. (2003) Inferring gene regulatory networks from time-ordered gene expression data of *Bacillus subtilis* using differential equations. *Pac. Symp. Biocomput.*, **8**, 17–28.
- De Hoon, M.J.L., Imoto, S., Kobayashi, K., Ogasawara, N. and Miyano, S. (2004) Predicting the operon structure of *Bacillus subtilis* using operon length, intergene distance, and gene expression information. *Pac. Symp. Biocomput.*, **9**, 276–287.

- Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.
- Friedman,N., Linial,M., Nachman,I. and Pe'er,D. (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.
- Hastie,T., Tibshirani,R. and Friedman,J. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York.
- Imoto,S., Goto,T. and Miyano,S. (2002) Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. *Pac. Symp. Biocomput.*, **7**, 175–186.
- Imoto,S., Kim,S.-Y., Goto,T., Miyano,S., Aburatani,S., Tashiro,K. and Kuhara,S. (2003) Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *J. Bioinform. Comput. Biol.*, **1**, 231–252.
- Kim,S.-Y., Imoto,S. and Miyano,S. (2003) Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *Comput. Meth. Syst. Biol., Springer-Verlag Lecture Notes in Computer Science*, **2602**, 104–113.
- Liang,S., Fuhrman,S. and Somogyi,R. (1998) REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.*, **3**, 18–29.
- Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, **6**, 127–138.
- Makita,Y., Nakao,M., Ogasawara,N. and Nakai,K. (2004) DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Res.*, **32**, D75–D77.
- Murphy,K. and Mian,S. (1999) Modelling gene expression data using Dynamic Bayesian Networks. *Technical Report*, Computer Science Division, University of California, Berkeley, CA.
- Ong,I.M., Glasner,J.D. and Page,D. (2002) Modelling regulatory pathways in *E.coli* from time series expression profiles. *Bioinformatics*, **18**(Suppl. 1), S241–S248.
- Segal,E., Yelensky,R. and Koller,D. (2003) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, **19**(Suppl. 1), i271–i282.
- Tamada,Y., Kim,S.-Y., Bannai,H., Imoto,S., Tashiro,K., Kuhara,S. and Miyano,S. (2003) Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, **19**(Suppl. 2), ii227–ii236.
- Van Someren,E.P., Wessels,L.F.A. and Reinders,M.J.T. (2000) Linear modeling of genetic networks from experimental data, *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, La Jolla, CA, August 2000. American Association for Artificial Intelligence, AAAI Press, Menlo Park, CA, pp. 355–366.
- Yada,T., Totoki,Y., Ishii,T. and Nakai,K. (1997) Functional prediction of *B.subtilis* genes from their regulatory sequences, *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, Halkidiki, Greece, June 1997. pp. 354–357.