BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–7

Bayesian sparse hidden components analysis for transcription regulation networks

Chiara Sabatti[1]*, Gareth M. James[2]

[1] Departments of Human Genetics and Statistics, UCLA, Los Angeles CA 90095-7088 and [2] Information and Operations Management Department, USC, Los Angeles, CA 90089-0809

ABSTRACT

Motivation: In systems like E. Coli, the abundance of sequence information, gene expression array studies, and small scale experiments allows one to reconstruct the regulatory network and to quantify the effects of transcription factors on gene expression. However, this goal can only be achieved if all information sources are used in concert.

Results: Our method integrates literature information, DNA sequences, and expression arrays. A set of relevant transcription factors is defined on the basis of literature. Sequence data is used to identify potential target genes and the results are used to define a prior distribution on the topology of the regulatory network. A Bayesian hidden component model for the expression array data allows us to identify which of the potential binding sites are actually used by the regulatory proteins in the studied cell conditions, the strength of their control, and their activation profile in a series of experiments. We apply our methodology to 35 expression studies in E. Coli with convincing results.

Availability: www.genetics.ucla.edu/labs/sabatti/software.html

Contact: csabatti@mednet.ucla.edu

1 INTRODUCTION

The complete sequencing of a large number of genomes, and the growing amount of information stored in databases allows us to identify genes, introns and exons, splice sites, binding sites for regulatory proteins, etc. As a consequence we can start tracing with some accuracy a picture of the possibilities inscribed in DNA sequences such as which proteins a cell could make, which transcription factors may regulate the expression of which genes, which alternative forms of a gene are possible. This complex collection of wiring systems has been described by Davidson (Davidson *et al.*, 2002) as a "view from the genome" of the cell. This static picture describes the realm of possibilities, rather than what actually happens in the cell.

Alternatively, one can talk about a "view from the nucleus", that offers a dynamic image capturing which genes are actually expressed, under the control of which transcription factor at any moment. Gene expression arrays, with all their limitations, by being a relatively low cost, high throughput experiment, conducted in a wide range of laboratories, offer a very important data source towards the gathering of such dynamic pictures. Indeed, there is a growing literature documenting attempts to reconstruct biological networks by applying statistical models to gene expression data. Many of these attempts are exploratory in nature, in that very little prior information on the structure of the network is assumed. While this line of work is clearly very important to help formulate hypotheses regarding yet unexplored mechanisms, in many cases enough information has been accumulated to enable us to take a more confirmatory approach. This paper describes such an approach with regard to the very specific process of transcription regulation, which is perhaps the first step linking the static information encoded in the genome with the dynamic system of the cell life. Figure 1 gives a schematic illustration of our approach: we reconstruct the structure and the dynamic behavior of the regulatory network involving a group of identified transcription factors (TF). We start with the analysis of the sequence of up-stream transcription units, obtaining an initial probability on the network topology. To then overcome the intrinsic limitation of a reconstruction based only on sequence data (which can provide, at best, a static description of the system), we turn to the analysis of a series of array experiments. We are able to refine our reconstruction of the network topology, as well as estimate changes in concentration of active form of TF, and the strength of their control of gene expression. To carry out the program illustrated in

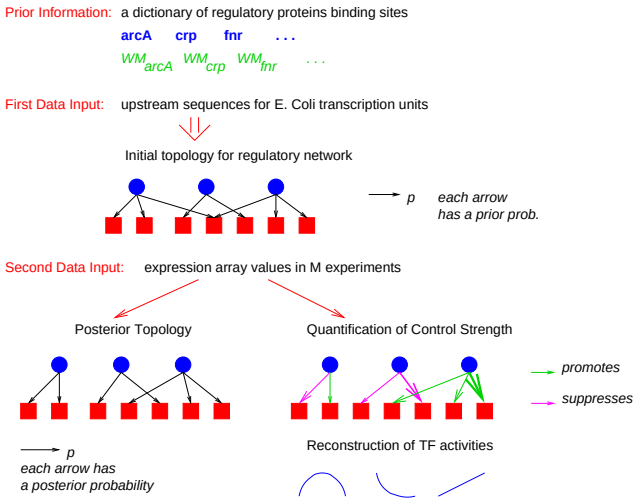

Fig. 1. Transcription network reconstruction integrating DNA sequence and gene expression information. Blue circles represent regulatory proteins and red squares genes. An arrow connecting a circle to a square indicates that the transcription factor controls the expression of the gene. When different colors are used in depicting these arrows, they signify a different qualitative effect of the TF on genes (repressor of enhancer). Finally, varying arrows thickness signify different control strengths.

*to whom correspondence should be addressed

Figure 1 we use a Bayesian framework to link sequence and array data and a hidden component model to relate gene expression values to the role of TF.

The paper is structured as follows. Section 2 describes how we use sequence analysis to obtain an initial probabilization of the network spaces. A model connecting expression values with information on binding sites is given in Section 3, while Section 4 gives details of our estimation strategy, from prior choices to design of the MCMC used in posterior exploration. Section 5 provides illustrations of the method on an E-coli data set. We conclude with a discussion.

2 PRIOR NETWORK TOPOLOGY

One of the goals of our analysis is the reconstruction of the network topology. The network of interest is defined starting from a set of transcription factors that are identified as important ones from the literature. We consider only simple two layer networks: parent nodes represent transcription factors, and descendent nodes regulated genes. Edges are directed and connect only TF to genes, so that we do not allow interactions between TFs or genes or feedback mechanisms. Such a simple network structure can be described with a 0-1 matrix Z, with N rows, as many as the genes under consideration, and L columns, where L is the number of transcription factors. The element z_{ij} is one if TF j regulates gene i, and zero otherwise. We deduce our initial distribution on the space of matrices Z from the analysis of the DNA sequence upstream of the studied genes. In particular, we use our *Vocabulon* (Sabatti and Lange (2002)) algorithm, as it is particularly well suited for this genomewide investigation, but the choice of the most appropriate methodology for this step is left to the investigators. Firstly, a group of transcription factors that are documented to have an important role in gene regulation of the system under study is selected. Prior available information on the characteristics of the DNA sequence motif they recognize informs the sequence analysis. We hence identify all the putative binding sites for these transcription factors in the portion of the genome sequence that is likely to have a regulatory function. The results of this initial analysis are used to define a prior distribution on the network topology as follows. We assume independence between the entries z_{ij}. Where there is documented experimental evidence of a binding site for transcription factor j in the promoter region of gene i, we set $z_{ij} = 1$ (for a detailed description of how this was obtained, please see Sabatti and Lange (2002)). Letting $\pi_{ij} = Pr(z_{ij} = 1)$, we assign a positive value smaller than 1 to the π_{ij} whenever the sequence analysis detects a putative binding site for transcription factor j upstream of gene i. The remaining entries of Z are set to zero. Note that one can use different thresholds to decide when a binding site is detected; moreover putative sites may have a varying degree of certainty that could be reflected in the choice of π_{ij}. In our experience, however, the most important issue is assuring that the prior is not excessively informative, allowing expression data to have a substantive contribution to the posterior. We suggest using the value $\pi_{ij} = 0.5$ for all the detected binding sites: this choice limits the informativeness of the prior distribution and we have found it to work well in practice.

The described prior distribution sets to zero the probability of a large number of edges. Overall, this adequately represents the nature of regulatory networks: the expression of every gene is affected only by a small number of transcription factors. Restricting a priori which edges are absent from the network results in computational advantages, greatly reducing the space of possible networks. However, note that this is equivalent to assuming that the binding site detection algorithm does not have false negatives. While this is clearly a limitation, it is important to recall that false negatives are not typically a serious problem for methods that identify the locations of binding sites of a known profile. Moreover, the threshold for detection can be lowered so to reduce the number of false negatives. Finally, note that configurations of Z with one or more rows containing only zeroes have positive probability according to the described distribution: the corresponding gene is excluded from the network.

3 GENE EXPRESSION DATA

To take advantage of the information contained in gene expression data, we need a model that links it to the network topology. For this purpose, we use a linear model that has been proposed in Liao *et al*. (2003) and Kao *et al*. (2004), and has elements of similarity with a number of other contributions such as Bussemaker *et al*. (2001), Keles *et al*. (2002), Conlon *et al*. (2003), Beal *et al*. (2005), and Girolami and Breitling (2004). The central assumption is that expression measurements can be thought of as determined by the, unknown, concentrations of active forms of the transcription factors. By log-transforming expression measurements, a linear model can be postulated, so that $e_{it} = \sum_{j=1}^{L} a_{ij} p_{jt} + \gamma_{it}$, where e_{it} represents the expression of gene i in experiment t; a_{ij} the control strength of transcription factor j on gene i; p_{jt} is a proxy for the concentration of active form of TF j in experiment t; and γ_{it} captures measurement errors and biological variability. We assume that γ_{it} are $i.i.d$ according to $N(0, \sigma_i^2)$ (we will indicate with σ^2 the vector of all these variance parameters, and with Σ a diagonal matrix with diagonal elements represented by σ^2.) It is useful to organize the model terms in matrices and vectors. For example, with E we indicate the matrix $\{e_{it}\}_{i=1,t=1}^{N,M}$, with M the total number of experiments analyzed. With e^t we indicate the t-th column of such matrix and with e_i the column vector corresponding to its i-th row. In matrix notation, the model we described can be expressed as:

$$E = AP + \Gamma, \tag{1}$$

where E represents the data and A and P unknowns. This formulation clearly underscores the fact that the model we are presenting is a factor analysis one (Anderson, 1984). Other hidden components models have been proposed in the literature for the analysis of gene expression data and we refer to the final section for a discussion of their relation to the present contribution. There we will also discuss the sense in which our linear model departs from others. Here it suffices to notice that (1) the "factors" p^j have a clear interpretation, as they correspond to specific transcription factors; (2) we are interested in reconstructing the specific values of the p_{jt}'s; (3) the matrix A is known to contain a large number of zeroes, corresponding to the sparsity of the network. These characteristics of the parameters inform our estimation procedure.

4 RECONSTRUCTION ALGORITHM

The first difficulty encountered in estimating a model like (1) is a lack of identifiability of A and P in a general setting. Given the interpretation of the parameters that we have described previously, a natural choice of constraints to achieve identifiability is restricting

the number of a_{ij} that are not equal to zero. The precise structure of these constraints necessary to achieve identifiability is described in Liao *et al*. (2003). Here we take a related, but substantially different approach. Liao *et al*. assume that the position of zeroes in A—corresponding to one regulatory network—is known a priori; the authors suggest checking if the defined model is identifiable and, in case of negative answer, they suggest eliminating some TF from the analysis. The present work does not assume a known regulatory network, but has as one of its main goals the reconstruction of one from sequence and array data. Furthermore, we want to be able to reconstruct a network, even if it does not satisfy the constraints in Liao *et al*. (2003), as they are not entirely biologically relevant. To carry out our program of network reconstruction and relax the identifiability requirements, we use a Bayesian framework. The sequence analysis described in Section 2 gives us a prior on the network structure, so that we are left needing to define the prior on $A, P, \sigma | Z$. While the initial distribution we defined for Z is based on biological information, and allows synergistic use of sequence and array data, the role of the prior on A, P, and σ is mainly one of regularization. This function can be performed while keeping the computational burden to a minimum by using conjugate-like priors. In this spirit, we consider each p_{jt} as a priori independent with a Gaussian distribution $p_{jt} \sim N(0, \sigma_p^2)$. The zero mean reflects the fact that a priori we do not know if the activity of the transcription factor j will be enhanced or reduced with respect to baseline in experiment t. The independence a priori and the common variance are useful for identifiability purposes and are related to the common assumption of identity matrix as variance-covariance for factors in frequentist factor models.

In a similar fashion, we assume that $a_{ij} = 0$ if $z_{ij} = 0$ and $a_{ij} \sim N(0, \sigma_a^2)$ otherwise, independently across i and j. The mean is set to zero as a priori one does not know if a transcription factor will act as a promoter or a repressor for a given gene. Note that choosing different values for σ_p^2 (and σ_a^2) one can obtain non-informative priors: we suggest an informative prior on P and a non informative on A (for a more complete discussion of prior choices, please see the supplementary material).

Note that the structure of some experiments may be such that p_{jt} and p_{js} are expected to be dependent (for example, t and s indicate two points in a time series). In such cases, it might be appropriate to assume a prior distribution $p^j \sim N(0, \Gamma)$; this will lead to a posterior distribution that is more complex than the one described in the following, but can nevertheless be explored with a Gibbs Sampler chain without substantially increased computational costs (see the supplementary material for the derivation).

Finally, we model σ_i^2, the variance of γ_{it}, as the inverse of a gamma distribution with parameters α_i and β_i. The value of the hyperparameters α_i and β_i can be determined using information derived from calibration slides or replicates of the array experiments. Indeed, often, the vector e_t is the average of the results of multiple replicate experiments in which case their variance can be adequately used to formulate a prior guess on the error variance.

In order to write out the relevant posterior densities of our parameters with compactness, we introduce some notation. If x and y are two r dimensional vectors, we denote by x^y the product of all the components of the first vector raised to the power of the corresponding components of the second i.e. $x^y = \prod_{i=1}^r x_i^{y_i}$. If z is a vector of zeros and ones, and a a vector of the same dimension, we indicate with $a[z]$ the vector of elements of a corresponding to ones in z. Similarly, if P is a matrix that has as many rows as z, $P[z]$ is the submatrix obtained by selecting the rows of P that correspond to ones in z. Moreover, if A has the same dimension as Z, A_Z indicates a matrix identical to A, except with all its elements corresponding to a zero in Z set to zero.

To explore the posterior distribution of the parameters, and perform inference, we use a Markov Chain Monte Carlo algorithm. Given the structure of our problem a collapsed Gibbs sampler is particularly convenient. There are four parameter groups Z, A, P and σ^2, and there is some conditional independence structure within each of these groups. The fact that we opted for conjugate-like priors guarantees that the full conditional distribution of the majority of the parameters have a known form, which makes it easier to run a Gibbs sampler. A detailed derivation of the algorithm can be found in the supplementary material. Here we describe the conditional distributions used in the algorithm:

$$\begin{aligned} \mathrm{P}(z^i | P, \sigma^2) \propto \pi^{i(z^i)}(1-\pi^i)^{(1-z^i)}/\sigma_a^{|z^i|} \times \\ \det(P[z^i]P[z^i]'/\sigma_i^2 + I_{|z^i|}/\sigma_a^2)^{-\frac{1}{2}} \times \\ \exp\{\frac{1}{2\sigma^4} e^{i'} P[z^i]'(\frac{P[z^i]P[z^i]'}{\sigma_i^2} + \frac{I_{|z^i|}}{\sigma_a^2})^{-1} P[z^i] e^i\} \end{aligned} \tag{2}$$

$$a_i | P, Z, \sigma^2 \sim N(\Sigma_{a_i} P[z_i] e^i / \sigma_i^2, \Sigma_{a_i}) \tag{3}$$

$$p_t | A, Z, \sigma^2 \sim N(\Sigma_{p_t} A_Z' \Sigma^{-1} e_t, \Sigma_{p_t}) \tag{4}$$

$$\frac{1}{\sigma_i^2} | A, Z, P \sim \mathrm{Gamma}(\tilde{\alpha}_i, \tilde{\beta}_i), \tag{5}$$

where I_r indicates an identity matrix of rank r, $\Sigma_{a_i} = (P[z^i]P[z^i]'/\sigma_i^2 + I_{|z^i|}/\sigma_a^2)^{-1}$, $\Sigma_{p_t} = (A_Z'\Sigma^{-1}A_Z + I_L/\sigma_p^2)^{-1}$, $\tilde{\alpha}_i = \alpha_i + M/2$, and $\tilde{\beta}_i = \beta_i + \sum_{t=1}^M (e_{it} - \sum_{j=1}^L a_{ij} p_{jt})^2/2$. One iteration of our algorithm consists of sampling z_i for $i = 1, \ldots N$ from (2), sampling a_i with $i = 1, \ldots N$ from (3), p_t with $t = 1, \ldots M$ from (4), and σ_i^2, $i = 1, \ldots N$ from (5).

The posterior probability (2) is unconditional on A (this is the collapsing step). Its value needs to be calculated for all possible z^i to sample from the appropriate multinomial probability. This is a potentially heavy computational burden. However, in general this will not be a problem because of the large number of zeros in Z i.e. for each gene, the number of potential binding sites is rather limited and this controls the dimension of the space of possible values of z^i.

The number of required iterations to reach convergence is typically unknown in MCMC on continuous state spaces like the present one. There are however a number of diagnostics that one can run on the chain in order to empirically assess if convergence has been achieved (see for example (Cowles and Carlin, 1995), and their implementation in the CODA R-package available from http://www-fis.iarc.fr/coda/). In our case this is slightly complicated by the fact that our posterior distribution is defined on a very high dimensional space. By monitoring autocorrelation of the chains, likelihood values, and using the Geweke and Hidelberger statistics we concluded that, with mildy informative priors, 11000 iterations with 1000 burn-in appear to guarantee convergence for the problem we describe (see supplementary material for details). The use of non informative priors makes the posterior multimodal and results in longer convergence times. Smaller size problems are likely to require fewer iterations; moreover, the chains seem to converge rather quickly to a region of high probability according to the posterior

and the large number of iterations is mainly necessary to explore the region around these quickly gathered modal values.

A significant advantage of our approach is the ease with which missing data in the expression matrix, E, can be handled. We simply add a fifth step to the Gibbs sampling algorithm described above where we impute any missing values. From (1) the distribution of $e_{it}|a^i, p_t$ is Gaussian with mean $\sum_{j=1}^{L} a_{ij}p_{jt}$ and variance σ_i^2. Hence, each iteration of the sampler is augmented to include sampling of e_{it}.

Once a sample from the posterior distribution is obtained, one can summarize it by calculating expected values and credibility intervals for each of the parameters. The combination of restrictions on A and or choices for the prior distributions, allow us to obtain a posterior distribution that is easy to deal with, except for an indeterminacy in the signs of A and P i.e. one can obtain identical results by flipping the sign on the jth column of A and the jth row of P. For the purpose of data analysis we defined the following quantities that are independent from rescaling and changes of signs and have interesting biological interpretations:

$$\tilde{p}_{jt} = \frac{\sum_i a_{ij}p_{jt}}{\sum_i 1(a_{ij}! = 0)} \quad \text{and} \quad \tilde{a}_{ij} = \frac{\sum_t a_{ij}p_{jt}}{M};$$

$\tilde{p}_{jt}$ is the average effect of each transcription factor on the genes it regulates (regulon expression), and $\tilde{a}_{ij}$ the average control strength over all experiments. These quantities are directly related to the expression values of genes in a regulon and for this reason we prefer them when conducting descriptive data analysis.

5 DATA ANALYSIS

We illustrate the applicability of our method with the analysis of 35 microarray experiments of E. Coli that are either publicly available or were carried out in the laboratory of Professor James C. Liao at UCLA. The experiments consist of Tryptophan timecourse data (1-12) (Khodursky *et al*., 2000), glucose acetate transition data (13-19) (Oh *et al*., 2000, 2002), UV exposure data (20-24) (Courcelle *et al*., 2001) and a protein overexpression timecourse dataset (25-35) (Oh and Liao, 2000). To reduce spurious effects due to the inhomogeneity of the data collection, we standardized the values of each experiment, so that the mean across all genes in each experiment is zero and the variance one. Merging these different datasets we have expression measurements on 4289 genes across 35 experiments. In general terms, biological knowledge of the nature of the microarray experiments suggests that the TrpR regulon should be activated in the Tryptophan timecourse, the LexA regulon should be activated in the UV experiments, and the RpoH regulon in the protein overexpression.

To define the network and our prior on the connectivity structure, we relied, as described previously, on literature knowledge and the results of a genomewide investigation for binding sites using a dictionary model (Sabatti *et al*., 2004). We categorized a location as a potential binding site if the Vocabulon algorithm assigned it a probability higher than 0.5. By merging these potential binding sites with the known sites from the literature, and with the expression data, we obtained a set of 1433 genes, potentially regulated by at least one of 37 transcription factors and on which expression measurements were available (missing values in the array data were allowed). Our prior on Z suggested a great deal of sparsity. For example, 14 of the transcription factors were only expected to regulate 20 or fewer genes and 34 of the 37 TFs were expected to regulate at most 120 genes. The notable exception was CRP, which potentially regulated over 500 genes. Let us note that without adopting our Bayesian framework, we would not be able to study this transcription network, simply because the number of experiments (35) is smaller than the number of TF considered (37): the use of priors regularizes the problem and enable us to see, a posteriori, that a number of the TF are not involved in this experiment.

The hyperparameters of the gamma distribution for error variances were all equal with $\alpha = .7$ and $\beta = .3$, leading to a weak prior. We experimented with the use of non-informative priors for A and P, by using $\sigma_a = 100$ and/or $\sigma_p = 100$. Not surprisingly, using an uninformative prior on both components leads to slow convergence and a multimodal posterior. When we regularized P by setting $\sigma_p = 1$ we obtained satisfactory and comparable results both when setting $\sigma_a = 1$ as well as when setting $\sigma_a = 100$. Using a non-informative prior on A, increases the number of a_{ij} evaluated as equal to 0 a posteriori, leading to simpler final models. All the analysis presented below is run with hyperparameter values $\sigma_p = 1$ and $\sigma_a = 100$.

The results from our analysis of the 35 experiments suggested that a significant portion of the potential binding sites should be discarded. For example, the posterior distribution on Z now contained 26 TFs that were expected to regulate 20 or fewer genes and 34 of the 37 TFs were expected to regulate at most 60 genes. Even CRP, went from over 500 potential binding sites in the prior to approximately 300 in the posterior. To better interpret the differences between the prior and posterior network, it is useful to underscore some characteristics of the process that lead us to the formulation of the prior. The search for binding sites carried out by Vocabulon is based uniquely on sequence information: it is quite possible that a portion of the E. Coli genome sequence looks just like a binding site for a TF, resulting in a high probability as estimated by our algorithm, but is actually not used by the protein in question. Moreover, the search for binding sites in the regulatory region of each gene is carried out by inspecting 600bp upstream of the start codon: given the size of E. Coli genes, this often results in investigating the same region for multiple (close together and short) genes. If a binding site is located in such a sequence portion, it will be recorded for all of the genes whose "transcription region" covers it. It is quite reasonable to assume that only one of the genes is actually regulated by the TF in question. In particular, one could decide in favor of the closest gene. However, such a choice is arbitrary. We have used the output of Vocabulon in a non-curated form for our prior, preferring to rely on array data to make such choices.

Figure 2 illustrates the regulon activities as reconstructed by our model: green dots indicate the posterior expected value and horizontal bars the confidence intervals for the activity levels of transcription factors in each experiment (experiments are organized along the vertical axis). The majority of the analyzed regulons are not perturbed by any of the experiments. This is to be expected, in that any shock induces a relatively small number of changes in the expression pathways. We repeated the analysis of the dataset, including only the transcription factors that appear to experience some changes in activation, and the genes that they regulate, and we obtained (for these TF) results entirely comparable to the ones shown here. This is not a surprise, given the sparsity of the connectivity, which makes it highly unlikely that one gene is regulated by more than one transcription factor. Another global observation

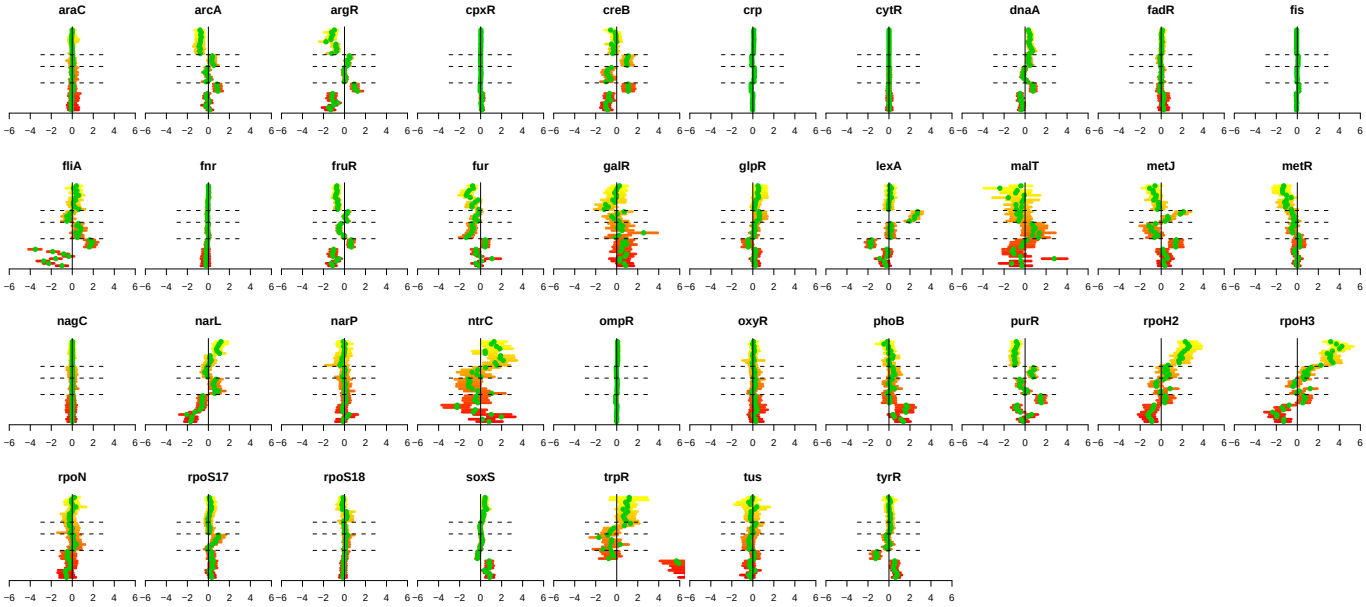

Fig. 2. Representation of regulons activities ($\tilde{p}^j$) for the 37 transcription factors in the study. Each graphical display refers to one transcription factor, whose name is reported on the top. Experiments are organized along the vertical axis, from bottom to top. Dashed horizontal lines separate the experiment groups for ease of identification. A green dot indicates, for each experiment, the posterior expected value of the transcription factor activity in that experiment. Horizontal bars provide posterior confidence intervals for the same parameters.

is that the location of the posterior distribution, and sometimes its spread, seems to vary across sets of experiments, even when the expected value of the regulon is not different from zero. This suggests that, despite our initial standardization, there may be residual differences in the noise levels of different experiments, which may be worth modeling. Additionally, one can notice the variability in the spread of the posterior distribution: this is inversely proportional to the number of genes attributed to the regulon. We conducted the same analysis using only binding sites documented in the literature (please see the appendix for details): in both cases we obtained wider confidence bands, underscoring the advantages of our methodology. Our results, instead, did not change appreciably when we changed our prior on Z to include a larger number of putative binding sites, defined by lowering the detection threshold. Most of the additional network edges where not supported by array data, leading to a reconstruction of regulon activities similar to the one described in Figure 2.

Focusing on the regulons that are activated in some of the experiments, we notice that our framework successfully brings to the attention of the researcher the regulons that are known to be affected by the type of shock experienced by the cell. We start looking at the first set of experiments, represented in the lower portion of the displays, from bottom up. The first 8 experiments (Khodursky *et al*., 2000) are two 4-point time courses of tryptophan starvation. The absence of tryptophan induces the de-repression of the genes regulated by trpR, and a clear increase in expression for this regulon can be observed. The experiments 9-12, instead, consider the effect of providing the cells with extra tryptophan, leading to opposite expectation for the trpR regulon: the posterior expected value is lower than zero, but the difference is not statistically significant. Additionally, it has been previously reported that addition of trpR downregulates several genes controlled by tyrR—and indeed, we notice a similar phenomenon. The patterns of argR and fliA regulon also correspond to previous literature observation (Khodursky *et al*., 2000). Figure 2 also suggests other effects (on the rpoH, narL, lexA regulon) that warrant further investigation. Experiments 20-24 are a comparison of wild type E. Coli cells with cells that were irradiated with ultraviolet light, which results in DNA damage. (The data points corresponding to this set are between the second and third horizontal dashed lines from the bottom of the displays.) Many of the DNA damaged-genes are known to be regularly repressed by lexA (Courcelle *et al*. (2001)). Indeed, according to our reconstruction, the lexA regulon experiences an increase in expression during these five experiments. Finally, we notice activation of a few regulons in the protein overexpression data. In particular, notice that rpoH2 and rpoH3 present the same profile across experiments (and increased expression in the last dataset): this is reassuring, since these two really represent the same protein, and are distinct here because they correspond to two different types of binding sites of the TF. Overall, it appears that our algorithm successfully captures the activation dynamics of the studied transcription factors. The fact that a considerable number of TF, however, do not seem to experience any change in the experiments, must significantly limit our ability to refine information on their binding sites and especially on the strength of their control.

As arguably the inference on the values of A is rather meaningless for those regulators that do not experience change in activity in any of the experiments, we focus our analysis on those which do. And for brevity, we give details only on the trpR regulon (some details on the lexA regulon are given in the supplementary material).

Figure 3 presents information on Z and $\tilde{a}$ for the trpR regulon. There were 4 genes known to be regulated by trpR and an additional

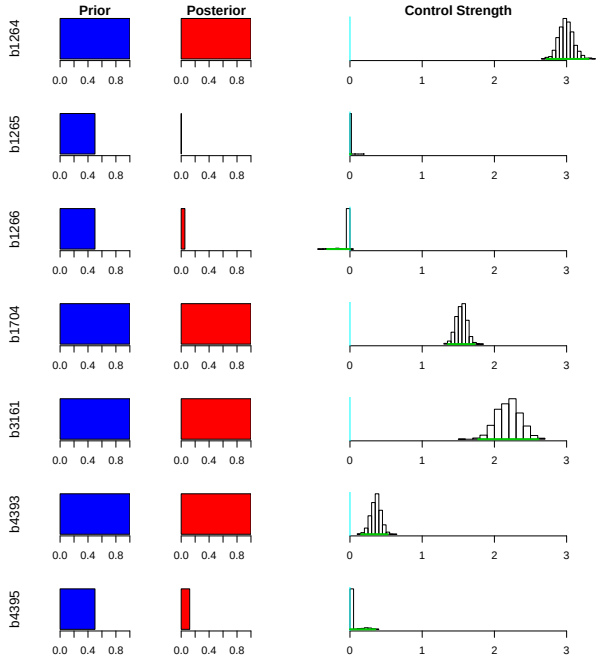

Fig. 3. The trpR regulon: connectivity and control strength information. Each row corresponds to one gene that can be potentially regulated by trpR. Genes are indicated by their "b-numbers." The first column represents the initial probability with which trpR is thought to regulate the target genes. The second column gives the corresponding posterior probability. The third column gives the histogram of sampled values of $\tilde{a}_{ij}$ for the considered gene.

3 imputed ones. Actually, the binding site suggesting the potential regulation of these three additional genes, is the same as that in the transcription region of two of the known genes, that is we have a couple of cases of the overlapping regulatory regions described above. The b-numbers, chosen to identify the genes, roughly correspond to their genomic location, so it is easy to see that the top three genes in the table are adjacent, and so are the bottom two. In the case of b1264, b1265, b1266, the last two genes appear to not be regulated by trpR: the posterior probabilities of a binding sites are very low. Thus it is possible to use our model to rule out the regulation of two genes by trpR, that are within a reasonable distance from a trpR real binding site. The case of the last gene in the list is similar. In summary, the analysis of the expression data helped us identify these three spurious binding sites, even ignoring that they were really multiple counts of the same site, which regulates only one gene. We would now like to point the reader's attention to the fact that genes b1264 (head of the trpR operon) and b4393–which are both known to be regulated by trpR, have very different a_{ij} values: this is reflected by the fact that in our dataset the first gene had much higher differential expression than the second one (see supplementary material) and different control strengths are needed to explain this variability.

To assess on a more general level the performance of our model and its ability to predict expression, as well as to compare it with a model that does not consider variable control strengths we conducted the following experiment. We randomly selected 12% of the expression values in E and set them to unknown (thus doubling the missing rate in our original dataset). We then ran our model on the remaining E values (training set) and used the estimated parameters to reconstruct the ones we blacked out (test set). We carried out the same experiment using a multivariate linear regression model for each experiment (as described in (Bussemaker *et al.*, 2001)), using as regressor our matrix of putative binding sites. The prediction error achieved with our model was half the prediction error obtained with the regression model. This is beacuse the regression model does not take into account variable control strength and does not evaluate posterior probabilities of existence of the binding sites. This suggests that the additional flexibility in our model is useful in capturing variation in gene expression.

6 DISCUSSION

While the number of studies that attempt reconstruction of gene networks from array data is large, the biological relations implied in these networks are very diverse. For example genes may be connected with an edge if they are coregulated, or if they belong to the same signaling or metabolic pathway, etc. We have focused on the much more specific domain of transcription regulation networks. Other contributions in this direction can be found in Liao *et al*. (2003) Beer and Tavazoie (2004), Segal *et al*. (2003), Gao *et al*. (2004), and Gardner *et al*. (2003). In these networks the activity of transcription factors is determined by the concentration of their active form, which depends largely on post-translational mechanisms. Hence, changes in mRNA levels for transcription factors are unlikely and are not necessary to cause substantial changes in their activity levels. Typically one has to augment the data on expression values with information on transcription factors derived from other sources (sequence analysis, ChIP-Chip data, experimental measurements on TF levels, literature knowledge, etc.) and/or model changes in the activity levels of transcription factors as hidden components. A few studies have been able to use measurements of transcription levels of regulatory proteins (see for example, Segal *et al*. (2003)); this strategy, however, is appropriate for only a relatively small fraction of transcription factors. For this reason, we assume that changes in TF activities are unobserved and we use sequence analysis to guide our reconstruction of these hidden factors.

One of the characteristics of our work is that a genomewide sequence analysis precedes and informs the interpretation of array experiments. Other studies start with the analysis of sequences by identifying a long list of putative regulatory elements and then refine these results by looking at expression values. In particular, Bussemaker *et al*. (2001), Keles *et al*. (2002), and Conlon *et al*. (2003) use a regression approach, which also resembles our linear model, to identify significant motifs, but their intentions differ substantially from ours. Their contributions aim to identify novel binding sites, not to quantify the extent of the control of a known regulatory protein on a gene. Additionally, they focus on the analysis of one array experiment.

We are not the first to use hidden components methodology to analyze gene expression data. Starting from Alter *et al*. (2000) there have been a number of applications of principal components or SVD to microarray data. The goals of these studies are mainly dimensionality reduction. There have also been a number of efforts to pursue more biologically minded analysis, using factor-like models. Perhaps the earliest work in this direction is West (2003), who suggests factor models to reduce the dimension of expression data to be used in linear models, paying particular attention to the development of

sparse models, in order to achieve a biologically realistic representation. Note that this same principle is reflected in our prior on Z. A very recent contribution is Girolami and Breitling (2004), where the authors focus on methods for factor analysis when distributions other then Gaussian appear appropriate. In Beal *et al*. (2005) a Bayesian version of state-space models is used to capture dynamical changes in gene expression in time series experiments as a function of unobserved biological changes, that can include activity levels of TFs. The substantial difference between our work and the approaches briefly quoted is that, in our scheme, hidden components correspond to identified TFs, instead of being objects defined purely on statistical grounds. Our results are immediately interpretable by referring back to the TFs that are represented in our study.

The contributions closest to ours are Beer and Tavazoie (2004) and Liao *et al*. (2003). Our contribution differs from Liao *et al*. (2003) in that we do not require complete prior knowledge of the network topology, but we reconstruct it in the course of the investigation. Moreover, adopting a Bayesian framework, we greatly relax the identifiability conditions as well as providing an easy mechanisim for evaluation of estimate variability. When the network topology is completely known our algorithm and the one in Liao *et al*. (2003) are equivalent. In contrast to the vast majority of the approaches outlined above, Beer and Tavazoie (2004) do not rely on a linear model for gene expression, but use a Bayesian network that allows them to capture non-linear effects. These authors' results suggest the presence of non-linear regulatory control and should motivate further investigation in this direction. Models like the one considered in this contribution, while not incorporating interaction effects between transcription factors, offer complementary information with respect to Beer and Tavazoie (2004). Consider, for example, how Beer and Tavazoie (2004) do not provide a prediction of the expression level for each gene in each experiment, but, for a given gene, simply predict to which cluster of expression profiles it belongs; clusters are defined with a preliminary data-analysis and based on correlation. This implies, for example, that differential behavior such as the ones of genes b1264 and b4393 described in the previous section cannot be accounted for. Furthermore, our model allows us to obtain a posterior probability of presence of binding sites for regulatory proteins in the gene up-stream regions, and to quantify different control strengths—an important departure from linear models such as, for example, the one described in Bussemaker *et al*. (2001).

ACKNOWLEDGMENTS

We thank professor James C. Liao and members of his laboratory (in particular Lars Rohlin) for providing us with expression arrays data sets. Chiara Sabatti was partially supported by NSF grants DMS0239427 and BES0120359, and NASA/Ames grant NCC2-1364.

REFERENCES

Alter O., P. Brown, D. Botstein (2000) Singular value decomposition for genome-wide expression data processing and modeling, *Proc Natl Acad Sci*, **97**, 10101–10106.

T. Anderson (1984) An introduction to multivariate statistical analysis, Wiley.

M.J. Beal, F. Falciani, Z. Ghahramani, C. Rangel, and D.L. Wild (2005) A Bayesian approach to reconstructing genetic regulatory networks with hidden factors *Bioinformatics*, **21**, 349–356

Beer, M. and S. Tavazoie (2004) Predicting gene expression from Sequence, *Cell*, **117**, 185–198.

rooks, S.P. and G.O. Roberts (1998) Assessing convergence of Markov Chain Monte Carlo algorithms, *Statistics and Computing*, **8**, 319–335.

Bussemaker, Li, Siggia (2001) Regulatory element detection using correlation with expression, *Nature Genetics*, **27**, 167–171.

Conlon, E., X. Liu, J Lieb, and J Liu (2003) Integrating regulatory motif discovery and genome-wide expression analysis *Proc Natl Acad Sci*, **100**, 3339–3344.

Courcelle, J., Khodursky, A., Peter, B., Brown, P.O. and Hanawalt, P.C. (2001) Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient Escherichia coli, *Genetics*, **158**, 41-64.

Cowles, MK and Carlin, BP (1995) Markov Chain Monte Carlo diagnostics: A comparative review, *J Amer Stat Soc*, **91**, 883–904.

Davidson et al. (2002) A Genomic Regulatory Network for Development, *Science*, **295**, 1669–1678

A. Dobra, B. Jones, C. Hans, J. R. Nevins and M. West (2004), Sparse graphical models for exploring gene expression data, *J. Mult. Analysis*, **90**, 196–212.

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

Gao, F., B. Foat, H. Bussemaker (2004) defining transcriptional networks through interactive modeling of mRNA expression and transcription factor binding data, *BMC Bioinformatics*, **5**, 31.

Girolami, M. and R. Breitling (2004) Biologically valid linear factor models of gene expression, *Bioinformatics*, **20**, 3021–3033.

K Kao, Y Yang, R Boscolo, C Sabatti, V Roychowdhury, and J Liao (2004) Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis, *Proc Natl Acad Sci*, **101**, 641-646;

Keles, van der Laan, and Eisen (2002) Identification of regulatory elements using a feature selection method, *Bionformatics*, **18**, 1167–1175.

Khodursky AB, Peter BJ, Cozzarelli NR, Botstein D, Brown PO, Yanofsky C. (2000) DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in Escherichia coli, *Proc Natl Acad Sci*, **97**, 12170–5.

C. E. Lawrence, S. F. Altschul, M. S. Bogouski, J. S. Liu, A. F. Neuwald, and J. C. Wooten (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment,*Science*, **262**, 208–214.

Liao, J., R. Boscolo, Y. Yang, L. Tran, C. Sabatti, and V. Roychowdhury (2003) Network component analysis: Reconstruction of regulatory signals in biological systems, *Proc Natl Acad Sci*, **100**, 15522–15527.

Oh, M.K., and J.C. Liao (2000) Gene Expression Profiling by DNA microarrays and Metabolic Fluxes in Escherichia coli, *Biotechnol. Prog.*, **16**, 278–286.

Oh, M.K., and J.C. Liao (2000) DNA Microarray Detection of Metabolic Responses to Protein Overproduction in Escherichia coli, *Metabolic Engineering*, **2**, 201–209.

Oh, M.-K., L. Rohlin, and J.C. Liao (2002) Global Expression Profiling of Acetate-grown Escherichia coli, *J. Biol.Chem.*, **277**, 13175–13183.

Quandt, K., K. Frech, H. Karas, E. Wingender, T. Werner, (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data, *Nucleic Acids Res.*, **23**, 4878–4884.

K. Robison, A. M. McGuire, and G. M. Church (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K12 genome, *Journal of Molecular Biology*, **284**, 241–254.

Roven, C., and H. Bussemaker (2003) REDUCE: an online tool for inferring *cis*-regulatory elements and trascriptional module activities from microarray data *Nucleic Acid Research*, **31**, 3487–3490.

C. Sabatti and K. Lange (2002) Genomewise motif identification using a dictionary model, *IEEE Proceedings*, **90**, 1803–1810.

Sabatti, C., L. Rohlin, K. Lange, and J. Liao (2005) Vocabulon: a dictionary model approach for reconstruction and localization of transcription factor binding sites, *Bioinformatics*, , .

Segal, E. et al. (2003) Module networks: identifying regulatory modules and their specific regulators from gene expression data, *Nature Genetics*, **34**, 166–76.

Tseng, G. C., M.-K. Oh, L. Rohlin, J. C. Liao, W. H. Wong (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects, *Nucleic Acids Research*, **29**, 2549–2557.

West, M. (2003) Bayesian factor regression models in the "Large p, Small n" paradigm, *Bayesian Statistics*, **7**, 723–732.

Yang, Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, T. P. Speed (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Research*, **4 e15**.