Title
Bayesian Sparse Hidden Components Analysis for Transcription Regulation Networks

Authors
Sabatti, Chiara
James, Gareth

Publication Date
2005-02-01

Bayesian sparse hidden components analysis for transcription regulation networks

Chiara Sabatti[1], Gareth James[2]

[1] Departments of Human Genetics and Statistics, UCLA, Los Angeles CA 90095-7088,

[2] Information and Operations Management Department, USC, Los Angeles, CA 90089-0809

UCLA Statistic Department Preprint # 414

February 2005

Running head Bayesian network component analysis.

Keywords Dictionary models; gene expression arrays; binding sites; transcription factors.

Corresponding author Chiara Sabatti

Department of Human Genetics

UCLA School of Medicine

695 Charles E. Young Drive South

Los Angeles, California 90095-7088 (USA)

FAX: (310) 794-5446

Phone: (310) 794-9567

e-mail: csabatti@mednet.ucla.edu

Abstract

We describe a framework where DNA sequence information and expression arrays data are used in concert to analyze the effects of a collection of regulatory proteins on genomic expression levels. The search for potential binding sites in sequence data leads to the identification of potential target genes for each transcription factor. The analysis of array data with a Bayesian hidden component model 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.

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 [8] 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. The knowledge derived from sequence analysis and experimentally verified binding sites for regulatory proteins, from which we can extract a static picture of cell regulation, can be taken as a starting point for further investigations that, with the addition of gene expression array data, aim to gather a dynamic, quantitative version of the same snapshot. 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.

To clarify our goal, we illustrate the transcription network with the graph in figure 1. Figure 1 contains a two layer network, where directed edges connect two types of nodes, parent nodes (indicated with R) represent transcription factors and descendant nodes represent genes whose expression is regulated by the ancestral transcription factors. A node can be either a parent or a descendent but not both. This excludes feedback processes that may be important, but are not necessary in a first order model of transcription regulation in simple systems such as E. Coli. A transcription factor will typically control multiple genes and each gene will be controlled by at least one transcription factor, and typically not more than three. The presence of an arrow connecting a transcription factor with a gene indicates the presence of a binding site for the transcription factor in the up-stream region of the gene, i.e. the potential role in regulation. The information represented in this network corresponds to the "view from the genome" picture we alluded to previously. We will refer to the information represented in figure 1 as topological: a list of nodes and directed edges. Aside from the topology of the network, there are other quantitative and dynamical characteristics of the biological system that one wishes to learn. For instance, a transcription factor influences genes with different degrees of strength. Hence, it is desirable to attach to each arrow in the network a numerical value signifying the control strength. Alternatively, while all arrows

correspond to possible binding sites, a given transcription factor will bind in front of a gene only in certain conditions and we want to be able to capture this dynamical behavior.

In this paper, we develop a Bayesian methodology to estimate the strengths of the effect of each transcription factor on each gene, the activation patterns of the transcription factors and the topology of the regulatory network. The Bayesian paradigm is particularly well suited to a problem of this nature because one can easily update prior information about the network topology, for example from experimentally verified binding sites, with new gene expression data to produce a new, more accurate, picture of the entire network. We illustrate this approach using E. Coli data, not only because it is a simple and well studied system, but also because it is widely used in biotechnical and bioengineering settings, so that the ability to simulate intracellular processes responds to a real practical need and its results will be tested by a large community of interested researchers. The choice of E. Coli as an initial organism of study also informs our choices of data sources. To define the topology of the network, we will rely substantially on sequence information—in the form of sequences upstream of the genes, where we will identify putative binding sites of regulatory proteins—and on database collections of experimentally identified binding sites for regulatory proteins of interest. There are other forms of high throughput experiments that can give some information on the location of binding sites, such as ChIP-Chip experiments. When such information is available, one would be wise to use it. However, such experiments are still rather rare and costly, and have not been carried out for E. Coli, which is why we decided to rely only on sequence data. To reconstruct the dynamical and quantitative aspects of the transcription regulation network we will rely on gene expression array experiments: a number of laboratories have now carried out large scale studies of this type for E. Coli and made their results publicly available, and this appears to be the case for a large number of organisms. The general flow of our procedure is illustrated in figure 2: we initially analyze sequence data to obtain an initial guess at the network topology and we then resort to array data to reconstruct activation profiles of transcription factors, their control strengths, and update the network topology.

The paper is structured as follows. Section 2 gives the basic model that we use to fit the network. A Bayesian framework is introduced in Section 3 to incorporate prior information we have about the network topology. In Section 4 we discuss the posterior distributions that this model induces and outline a Gibbs sampling procedure for fitting the model. Section 5 provides illustrations of the method on an E-coli data set. We conclude with a discussion.

2 A model for gene expression data

Consider N genes that are known to be regulated (over all) by L transcription factors. Suppose that the expression values of the genes can be measured with gene expression arrays, leading to a vector of background corrected, normalized and log transformed values $e = (e_1, \ldots, e_N)'$ (see [21, 29]) Note that the terminology we use is typical of cDNA spotted arrays, but the analysis is unchanged when a different experimental platform is used. Furthermore, if cDNA arrays are used, typically e will represent changes from a baseline value of expression and all the quantitative values in the network will need to be interpreted in relation to changes from this baseline. For simplicity we will refer to e as the expression level, an absolute quantity.

We want to relate gene expressions to quantitative values representing the activity of transcription factors. Since their concentration in the nucleus is crucial in defining how often they bind to recognized DNA sites, we let p, a $L \times 1$ vector, represent the concentrations of active forms of the transcription factors. Let A be a $N \times L$ matrix, whose elements a_{ij} quantify the control strength of transcription factor j on gene i, where $a_{ij} = 0$ signifies no relationship. Suppose that we perform M different experiments. Then our model can be written as

$$e_t = Ap_t + \gamma_t, \quad \gamma_t \sim \mathcal{N}(0, \sigma^2 I), \quad t = 1, \ldots, M \tag{1}$$

with e_t and γ_t respectively representing the observed expression levels and an error term for experiment t, and p_t and A unknown quantities. Our model poses a linear relation between e_t, the

gene expression values, and p_t, the concentration of active forms of transcription factors. We are not assuming that such a relation corresponds to the biological reactions underlying transcription regulation, but simply suggesting it as a viable approximation for the purpose of our data analysis. Indeed, there are a number of contributions in the literature where such a simple model appears to be useful in capturing at least first order effects ([5, 14, 6]). A crucial component of our linear formulation is that it excludes the presence of interactions between transcription factors. While this hypothesis may be too restrictive in general, it is quite adequate for E. Coli, where there are very few examples of known interactions between transcription factors. Note that the model presented in (1) is identical to the one adopted in [17] and [13] and called network component analysis. These articles also provide motivation for the biological grounding of this simple bilinear model.

We further assume that the error terms in γ_t be independent, with a Gaussian distribution and zero mean. The error distribution clearly depends on the quality of the data. As we mentioned before, we assume that the values e_t have been pre-processed with state of the art techniques that correspond to the specific array platform used which insure that the measurement error is as close to white noise as possible. Independence is not a very strong assumption, if one believes that the linear terms in the model have captured the dependence between genes, as it is due to regulation by common factors. The choice of a Gaussian distribution is more arbitrary, but does provide significant reductions in computational difficulty and, in addition, once we deal with estimation, leads to the same least squares criteria that was adopted in [17].

In a more general setting, models with the form given in (1) are often called factor analysis models [3]. There are, however, some aspects that make our problem quite different from the standard factor analysis. First, in our setting the factors p are considered parameters and not random quantities: this is because in different experiments one expects different and unique concentration values that depend on the cell enviroment. Secondly, often, in factor analysis an a-priori interpretation of the factors is not available, while in our setting, each factor in the model corresponds to one specific regulatory protein. A third departure from the most common venue of factor analysis

depends on how we approach the problem of identifiability.

Let us comment on the nature of the unknowns, p_t and A. The regulation of transcription factors mainly occurs post-translationally. In response to varying cell enviroments, these molecules undergo chemical modifications which transform them between active and inactive forms. These modifications are substantially faster than the synthesis of new proteins and hence enable the cell to efficiently respond to stimuli. The fact that these changes happen post-translationally implies that measuraments of the mRNA transcripts of these genes do not provide relevant information. For instance, in E. Coli the vast majority of transcription factors are expressed by the cell at a constant rate over time. In some other organisms, such as yeast, the transcription factors involved in cell cycles appear to undergo changes in expression level [27]. Nevertheless, this is the exception rather than the rule. As the expression levels of transcription factors are typically constant and not directly related to their activity, the data e_t does not provide direct information on the concentration of the active form of transcription factors. Additionally, direct measurements of these concentrations are very difficult, given their typically very low levels. For these reasons, we consider p_t an unknown parameter.

The matrix A contains information on the arrows, i.e. the topology, in the transcription network. A value of $a_{ij} = 0$ implies that there is no arrow, or equivalently no relationship, between transcription factor j and gene i. Other numerical values of a_{ij} indicate the control strength of transcription factor j on gene i. If no information is available for network topology then one is forced to perform a blind deconvolution of e_t to estimate the two unknown quantities A and p_t. Typically, this will require using methods such as principal components analysis which make strong statistical assumptions about p_t and A (see [2], for example). Alternatively, if A is completely known, then the problem of estimating p_t based on the expression levels, e_t, can be performed using standard least squares. However, in reality we have partial information about A and we find ourselves somewhere between these two scenarios. For instance, because of the topological characteristics of our network, we know that the majority of the elements of A will be zero. In addition, we

can identify putative binding sites of regulatory proteins (potential non-zero elements of A) using sequence information, in the form of sequences upstream of the genes and database collections of experimentally identified binding sites for regulatory proteins of interest. Hence, we adopt a Bayesian framework which allows us to incorporate this partial information into the model given by (1) in the form of priors on A and p_t. We discuss these priors in the following section.

3 Prior distribution

An essential characteristic of A is that it contains a large number of zeroes so our prior on A has to incorporate this knowledge. For this purpose, we define a matrix Z with the same dimensions as A such that each element, z_{ij}, is either zero or one. Values of $z_{ij} = 0$ imply $a_{ij} = 0$ and hence no relationship between gene i and transcription factor j. Values of $z_{ij} = 1$ imply that $a_{ij} \neq 0$. In this formulation Z defines the network topology and a_{ij} the strength with which transcription factor j acts on gene i. For each element of Z we assign a probability of a connection, $\pi_{ij} = Pr(z_{ij} = 1)$, and we assume independence across the z_{ij}'s. Where there is documented experimental evidence of a binding site for transcription factor j in the promoter region of gene i, we set $\pi_{ij} = 1$. For the remaining elements of Z we calculate π_{ij} based on an analysis of the sequence upstream of the gene under consideration. While this approach can be taken with a variety of methods to identify binding sites, we used our own *Vocabulon* algorithm ([26]) that assumes a dictionary model with variable spelling and is precisely apt to scan long sequences for the occurrence of multiple motifs. Vocabulon identifies, for each transcription factor in the dictionary, a series of putative locations and evaluates, for each of these, a probability of occurrence. Technically, a probability of occurrence is evaluated for each binding site at each location but a user discretional threshold can be used to select locations that have a substantial probability. In our case we chose to set π_{ij} equal to the predicted probability for estimated values larger than 0.05 and $\pi_{ij} = 0$ for lower values. This strategy assumes that the Vocabulon algorithm contains no false negatives. This

assumption is certainly a limitation, but not a serious one given that we can fix the threshold for detection as low as desired and the fact that in general false positive are a much more serious problem than false negatives for these algorithms. This choice greatly reduces the set of possible non-zero a_{ij}'s and hence the computational burden associated with their exploration.

Given z_{ij}, the prior probability on a_{ij} is simply set by letting

$$a_{ij} = \begin{cases} \mathcal{N}(0, \sigma_a^2) & z_{ij} = 1 \\ 0 & z_{ij} = 0, \end{cases}$$

independently across i and j. The choice of a Gaussian distribution is dictated by convenience. Its mean is set to zero as a priori one does not know if a transcription factor would act as a promoter or a repressor of a given gene. The constant variance acts as a regularization parameter.

The prior distribution we choose for the p_t parameters is very similar to the one of the non zero components of the matrix A. We assume L transcription factors observed over M different experiments. Hence we can form a $L \times M$ matrix P of parameters. Our prior takes each p_{jt} as a priori independent with a Gaussian distribution $p_{jt} \sim \mathcal{N}(0, \sigma_p^2)$. Again, Gaussianity is chosen for computational convenience. 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 and the common variance and independence a priori are useful for identifiability purposes. Note that assuming that factors are independent a-priori is not unrelated to the assumption of independence that is common in factor analysis models where factors are often considered random variables. Both contribute to identifiability but when, as in our case, factors are modeled as parameters, their independence is merely in the prior distribution, with their posterior most likely incorporating some dependence whereas when factors are modeled as random variables, independence remains one of their structural characteristics.

Finally, we model σ^2, the variance of γ_t, as the inverse of a gamma distribution with parameters α and β. Note that assuming the same error variance for all genes may be unrealistic and is

indeed unnecessary. We do so purely for notational convenience, as none of the formulas we will derive changes substantially if we assume a different variance for each gene. The value of the hyperparameters α and β 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.

Combining these priors with the model from (1) we obtain:

$$e_t = Ap_t + \gamma_t, \quad \gamma_t \sim \mathcal{N}(0, \sigma^2 I), \quad t = 1, \ldots, M,$$

with priors on the parameters a_{ij}, p_{jt} and σ^2 of

$$a_{ij}|z_{ij} = 1 \sim \mathcal{N}(0, \sigma_a^2), \quad p_{jt} \sim \mathcal{N}(0, \sigma_p^2), \quad Pr(z_{ij} = 1) = \pi_{ij}, \quad \frac{1}{\sigma^2} \sim Gamma(\alpha, \beta). \quad (2)$$

In this formulation $\sigma_a^2, \sigma_p^2, \pi_{ij}, \alpha$ and β are considered hyperparameters. It is useful to compare this model to that used in the studies [17, 13]. There, a model similar to (1) was used except no priors were assumed for A, p_t or σ^2 and the z_{ij}'s were taken to be known parameters. In other words the network topology was supposed to be completely specified with only the strength of the effect of transcription factors unknown. Such an assumption is unrealistic in practice because, while some transcription factor-gene relationships are well documented, most are only hypothesised relationships based on methods such as the Vocabulon algorithm. The Bayesian approach using the priors given by (2) has two advantages. First, by placing a prior on the z_{ij}'s, or equivalently the network topology, we can incorporate partial information without assuming full knowledge of the network. Second, through the generation of posterior distributions, described in the following section, one can easily produce estimates of uncertainty associated with each of the model parameters.

4 Posterior distribution and inference

The central instrument of inference in our Bayesian model is the posterior distribution of the parameters, Z, A, P and σ^2. Once we evaluate the posterior distribution we can produce point estimates, using the posterior means, and confidence intervals, using the appropriate posterior quantiles. In Section 4.1 we derive the posterior distribution. Section 4.2 shows how we use a Gibbs sampler to sample from the posterior distribution and Section 4.3 discusses identifiability issues.

4.1 The posterior distribution

In order to write out the posterior density of our parameters with some 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 matrix, we denote the vector corresponding to its tth column by z_t, and the column vector corresponding to its ith row by z^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.

Then, since A, P and $E|Z, A, P$ all have a Gaussian distribution the posterior can be written

as

$$
\begin{aligned}
Pr(Z, A, P, \sigma^2|E) &\propto Pr(Z, A, P, \sigma^2, E) = Pr(E|Z, A, P, \sigma^2)Pr(\sigma^2, Z, A, P) \quad (3)\\
&= Pr(E|Z, A, P, \sigma^2)Pr(\sigma^2)Pr(Z)Pr(A|Z)Pr(P)\\
&\propto \left(\frac{1}{\sigma^2}\right)^{\frac{MN}{2}+\alpha-1} \exp\left\{-\frac{1}{2\sigma^2}\left(2\beta + \sum_{i=1}^{N}(e^i - P[z^i]'a^i[z^i])'(e^i - P[z^i]'a^i[z^i])\right)\right\} \times\\
&\quad \left[\prod_{i=1}^{N} \pi^{i(z^i)}(1-\pi^i)^{(1-z^i)}\right] \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{N} a^i[z^i]'a^i[z^i]/\sigma_a^2 + \sum_{t=1}^{M} p_t'p_t/\sigma_p^2\right)\right\} \quad (4)
\end{aligned}
$$

Notice that the form of the likelihood function, $Pr(E|Z, A, P, \sigma^2)$, derives from the fact that the sum that appears in the exponential of the likelihood can be written in a few alternative forms:

$$\sum_{i=1}^{N}\sum_{t=1}^{M}(e_{it} - \sum_{j=1}^{L} a_{ij}p_{jt})^2 = \sum_{i=1}^{N}(e^i - P[z^i]'a^i[z^i])'(e^i - P[z^i]'a^i[z^i]) = \sum_{t=1}^{M}(e_t - A_Z p_t)'(e_t - A_Z p_t).$$

4.2 Exploration of the posterior and inference

In order to sample from, and hence estimate, the posterior distribution, it is convenient to use a collapsed Gibbs sampler, which is an example of a Markov chain. A Gibbs sampler works by iteratively producing a random sample from one of the parameters conditional on the previously sampled values of the other parameters. In order to implement such an algorithm we must identify the conditional distributions of the four parameter groups Z, A, P and σ^2. Notice, firstly, that a^i is independent from a^j for $i \neq j$ conditionally on Z and P. Similarly, p_t is independent from p_s conditionally on A and Z. Then the four conditional distributions derived from the posterior

distribution, given by (4), are:

$$\mathrm{P}(z^i|P,\sigma^2) \propto \pi^{i(z^i)}(1-\pi^i)^{(1-z^i)} \times \frac{\det(P[z^i]P[z^i]'/\sigma^2 + I_{|z^i|}/\sigma_a^2)^{-\frac{1}{2}}}{\sigma_a^{|z^i|}} \exp\{\frac{1}{2\sigma^4} e^{i'} P[z^i]'(P[z^i]P[z^i]'/\sigma^2 + I_{|z^i|}/\sigma_a^2)^{-1} P[z^i]e^i\} \quad (5)$$

$$a_i|P,Z,\sigma^2 \sim \mathcal{N}((P[z^i]P[z^i]'/\sigma^2 + I_{|z^i|}/\sigma_a^2)^{-1}P[z_i]e^i/\sigma^2, (P[z^i]P[z^i]'/\sigma^2 + I_{|z^i|}/\sigma_a^2)^{-1}) \quad (6)$$

$$p_t|A,Z,\sigma^2 \sim \mathcal{N}((A_Z'A_Z/\sigma^2 + I_L/\sigma_p^2)^{-1}A_Z'e_t/\sigma^2, (A_Z'A_Z/\sigma^2 + I_L/\sigma_p^2)^{-1}) \quad (7)$$

$$\frac{1}{\sigma^2}|A,Z,P \sim \text{Gamma}(\alpha + MN/2, \beta + \sum_{i=}^{N}\sum_{t=1}^{M}(e_{it} - \sum_{j=1}^{L} a_{ij}p_{jt})^2/2). \quad (8)$$

The conditional distributions given by (6), (7) and (8) are easily identified from the joint posterior. The posterior probability for each vector z^i conditional on P and σ^2, (5), is obtained by integrating out a^i and only keeping track of the terms that depend on z^i. This is the collapsing step of the Gibbs sampler and it is useful to resort to it in order to maximize the mixing of the chain. Note that we need to calculate (5) for all possible z^i and then sample according to a 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 hence so to is the space of possible values of z^i. Once a sample from the posterior distribution is obtained, one can summarize it by calculating expected values and confidence intervals for each of the parameters.

Another 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^2. Hence, at each iteration of the sampler, we impute e_{it} using its conditional mean, where a^i and p_t are obtained from the most recent Gibbs sample.

Furthermore, note that in some cases one may not want to assume independence a priori among p_{jt} and p_{js} as t and s are experiments in a time series. To describe cases of this generality, we can

assume a prior distribution $p^j \sim \mathcal{N}(0, \Lambda)$; the posterior distribution of p, then, cannot be separated in the independent p_t components, but can nevertheless be explored with a Gibbs Sampler chain.

4.3 Identifiability

There is a fundamental indeterminacy in any factor analysis type model such as the one we propose in (1). Indeed, if one takes any invertible $L \times L$ matrix, and defines $\tilde{A} = AX$ and $\tilde{P} = X^{-1}P$, one obtains a new set of parameters that lead to exactly the same likelihood for any set of observations since

$$\tilde{A}\tilde{P} = AXX^{-1}P = AP. \tag{9}$$

In classical statistics such models are called unidentifiable because it is not possible to discriminate, on the basis of the data, among a class of possible parameter values. In factor analysis a series of restrictions are imposed on A and P to overcome this impasse [3]. The nature of these constraints depends on the specific type of factor model considered: for example, when factors are taken as random variables with a Gaussian distribution, one can impose constraints on the variance-covariance matrix. When factors are considered parameters, other options are available. For example, in [3] it is shown that identifiability can be achieved by constraining certain elements of the A matrix to be zero. A sparsity constraint of this form is biologically sensible, as we know that genes are generally influenced by a small number of transcription factors, and we have considerable prior knowledge to identify these zero elements. Indeed, [17] provides a general set of conditions on the sparsity of A under which the only matrices X for which (9) holds are diagonal and hence A and P are identifiable up to multiplicative constants. Unfortunately, while the constraints in [17] are overall rather reasonable, it is unrealistic to assume that the true transcription network will respect all the identifiability constraints given in [17].

In this paper, we take a Bayesian approach, which has implications with regard to identifiability. Generally speaking, the fact that multiple parameter values are equally favored by the data

does not represent a radical difficulty for Bayesian inference, as long as the posterior distribution is proper (it is still possible, for example, to calculate the posterior expected value). Nevertheless, unidentifiable models may lead to posterior distributions that have multiple modes of equal value, which represent an impasse if we are interested in maximum a posteriori estimates and also creates computational difficulties when we need to explore the posterior distribution with MCMC methods. For these reasons, it is preferable to choose the prior distribution so that the posterior is unimodal.

In our case, the combination of restrictions on A and 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. To deal with this sign indeterminacy, one can adopt a series of possible conventions. We opted to constrain the mean value for each row of P to be positive. This constraint was achieved using a two step approach. First, for any iteration of the Gibbs sampler where a row of P had a negative mean we flipped the signs on the corresponding row of P and column of A. This approach has the potential disadvantage of incorrectly flipping the sign simply because random fluctuation in a Gibbs iteration caused a sample row mean to be negative even though the true row mean was positive or vice versa. Hence, for the second step we computed $\bar{P}_j$, the mean value of the jth row of P over all the Gibbs samples. We then calculated the mean squared difference both between P_j and $\bar{P}_j$ and between $-P_j$ and $\bar{P}_j$ for each Gibbs iteration. If $-P_j$ was closer to $\bar{P}_j$ we again flipped the sign. In addition we standardize each row of P to have norm one and correspondingly adjust the columns of A. In theory, this normalization is not required because of the priors on A and P. However, we have found in simulations that the estimates are far more stable when normalized, especially when the priors are weak.

In addition to the direct estimates for A and P we also compute the average effect of each transcription factor on the genes it regulates (regulon expression) $\tilde{p}_{jt}$, and the average control

strength over all experiments, $\tilde{a}_{ij}$. In particular

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

These quantities are more directly related to the expression values of genes in a regulon and for this reason we prefer them when conducting descriptive data analysis. The normalized values of A and P described above may be more useful for prediction purposes.

5 Data analysis

We illustrate the applicability of our method with the analysis of 35 microarray experiments of E. Coli that are either publically available or were carried out in the laboratory of Professor James C. Liao at UCLA. The experiments consist of Tryptophan timecourse data (1-12) [15], glucose acetate transition data from the Liao lab (13-19) [18, 19], UV exposure data (20-24) [7] and a protein overexpression timecourse data also from the Liao lab (25-35) [20]. To reduce spurious effects due to the inhomogeneity of the data collection, we standardized the values of each experiments, 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 LexA regulon should be activated in the UV experiments, the TrpR regulon should be activated in the Tryptophan timecourse, 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. For details on the latter we refer to the original paper [26]. 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). The top portion of Figure 3 illustrates the level of sparsity implied in the prior distribution. For example, our prior suggests that 14 of the 37 transcription factors each regulate at most 20 genes and that approximately 1000 of the genes, the vast majority, are regulated by only 1 transcription factor. Each transcription factor is expected to regulate a rather small number of genes, with the notable exception of CRP, which potentially regulates over 500 genes.

Figures 3 through 5 give a global view of the results from our analysis of the 35 experiments. Figure 3 (lower portion) illustrates how the analysis of the array data modifies our prior belief in the network structure: a significant portion of the potential binding sites are discarded. To better interpret the differences between the prior and posterior network, it is useful to underscore some characteristics of the process that leads 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 inspecting 600bp upstream 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 are 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. In order, however, for this to be possible, given the relatively small size of the array data set, we had to down-weight the probabilities calculated by Vocabulon, which will be uniformly high in case of the overlap described above. In particular, we have, quite arbitrarly, set equal to 0.5 the probability of each binding sited detected by Vocabulon,

but not known to be true in the literature.

Figure 4 illustrates the regulon activities as reconstructed by our model: green dots indicate the expected value and the vertical bars span the regions that receive 0.99 probability according to the posterior distribution. The first piece of information quickly conveyed by Figure 4 is that 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 relative 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. An other global observation is that the location of the posterior distribution, and sometimes its spread, seem 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, note that not surprisingly the spread of the posterior distribution is inversely proportional to the number of genes in the regulon.

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. In particular, the first 8 experiments [15] 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 [15]. Figure 4 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. Many of the DNA damage-genes are known to be regularly repressed by lexA. 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, hence, 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.

Figure 5 gives an overall image of our results in estimating control strengths. While it is difficult, and arguably not too meaningful, to extract general patterns, one can notice that a large portion of the confidence intervals for $\tilde{a}$ cover zero, as one might expect due to the lack of information about the TF's that do not experience changes in activation in the set of considered experiments. It is more relevant to discuss the case of the activated regulons. We focus on trpR and lexA. Figure 6 presents information on Z and $\tilde{a}$ for the trpR regulon. There were 4 genes known to be regulated by trpR and an additional 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. b1265 can be excluded by at the posterior probability of regulation. b1266 has posterior

probability higher than 0.5, but its control strength is not significantly different from zero and is in the opposite direction to that of the four genes known to be regulated by trpR—which is known to act only as a repressor. Thus it was 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, however this time both the posterior probability of regulation is high and the majority of the sampled $\tilde{a}$ values agree with the ones of the close-by gene, truly regulated by trpR. We can either hypothetize that these genes form a weak operon, or that there are some errors in transcription so that b4359 is often transcribed when b4393 is.

Figure 7 gives some details of the analysis of the lexA operon (that contains 45 genes). The first two genes, b0958 and b0959, adjacent, with b0958 truly regulated by lexA, and the second hypothetized to, because its regulatory region contains the lexA binding site for b0958. A look at the posterior values for Z and $\tilde{a}$ clearly rules out the possibility that b0959 is regulated by lexA. The last group of genes represent a similar situation, but this time array analyis leadsone to believe that both genes are regulated. Indeed, the latest version of the database regulonDB documents this as a potential operon.

6 Discussion

The literature on gene networks and their reconstruction using microarray expression experiments contains so many articles that it is almost impossible to review them entirely and contrast them with the specific approach reported here. Hence, we will utlize some overarching themes to organize the discussion, with the very specific goal of identifying the contributions that are closest to ours, clarifying its originality.

Firstly, we want to point out that 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 be-

long 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 [17],[4],[27],[11], [10]. One of the fundamental characteristics of transcription regulation networks is that the activity of transcription factors is determined by the concentration of their active form, which depends largely on post-translational mechanisms. In other words, changes in mRNA levels for transcription factors are unlikely and are not necessary to cause substantial changes in their activity levels. This implies that 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. Few studies have been able to use measurements of transcription levels of regulatory proteins (see for example, [27]); this strategy, however, is appropriate for only a relatively small fraction of transcription factors, typically cell-cycle related. We assume that changes in TF activities are unobserved and we use sequence analysis to guide our reconstruction of these hidden factors. We now briefly consider other contributions that share similar premises.

Sequence and expression array information have been previously used in concert. Indeed, there are a large number of cases where a novel regulatory motif is discovered after analysis of the upstream sequences of genes that exhibit a common regulatory mechanism. In such cases, array analysis precedes sequence analysis, which is the opposite order to that of our approach. A few 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, [5, 14, 6] 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 by no means the first to use hidden components methodology to analyze gene expres-

sion data. Starting from [2] 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 [28], 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 [12], where the authors focus on different distributional assumptions. In [1] 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 transcription factors. Our work differs substantially from others that use hidden components as in our case these, while unobserved, are specifically identified with known transcription factors through the use of prior knowledge. This means our approach is uniquely tailored to this problem, particularly powerful, and produces easy to interpret results.

Perhaps the contributions closest to ours are [4] and [17]. The methodology described in [4] identifies possibly relevant sequence elements from the analysis of upstream regions of genes that appear to have similar expression behavior. No prior information on regulatory proteins and the form of their binding sites is assumed. On the contrary, one of our aims is precisely to quantify the role of known transcription factors. Furthermore, the relation between sequence elements and gene expression is modeled in [4] using a Bayesian network, while we propose a simple factor model. A factor model is also adopted in [17]. However, our contribution differs from [17] in that we adopt a Bayesian framework, which greatly relaxes the identifiability conditions as well as providing an easy mechanisim for the inclusion of partial information about the network topology. We do not require absolute knowledge on the position of binding sites from the literature. Instead we analyze sequence data with the Vocabulon algorithm to gather prior information on possible sites. Out method also has little difficulty dealing with expression array data containing missing

information.

A limitation of our current implementation is in the assumption of an absence of false negative results in our sequence analysis for identification of binding sites. This would require modifying our prior probability so that any regulatory protein has a non zero prior of having a binding site in front of any gene, while ensuring sparsity. The approach of [9] may provide some suggestions in this regard. Concrete implementation is likely to considerably increase the computational burden.

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 ASA/Ames grant NCC2-1364.

References

[1] 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

[2] 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.

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

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

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

[6] Conlon, E., X. Liu, J Lieb, and J Liu Integrating regulatory motif discovery and genome-wide expression analysis PNAS 2003 100: 3339–3344.

[7] 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.

[8] Davidson et al. (2002) A Genomic Regulatory Network for Development. Science 295: 1669-1678

[9] 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.

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

[11] 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.

[12] Girolami, M. and R. Breitling (2004) Biologically valid linear factor models of gene expression, Bioinformatics, *in press*.

[13] K Kao, Y Yang, R Boscolo, C Sabatti, V Roychowdhury, and J Liao Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis PNAS 2004 101: 641-646;

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

[15] Khodursky AB, Peter BJ, Cozzarelli NR, Botstein D, Brown PO, Yanofsky C. DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in Escherichia coli. Proc Natl Acad Sci U S A. 2000 Oct 24;97(22):12170-5.

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

[17] Liao, J., R. Boscolo, Y. Yang, L. Tran, C. Sabatti, and V. Roychowdhury (2003) "Network component analysis: Reconstruction of regulatory signals in biological systems" *PNAS* **100**: 15522–15527.

[18] 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.

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

[20] 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.

[21] Tseng, G. C., M.-K. Oh, L. Rohlin, J. C. Liao, W. H. Wong. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Research, 2001, Vol. 29, No. 12 2549-2557.

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

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

[24] 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*, vol. 31: 3487–3490.

[25] C. Sabatti and K. Lange, “Genomewise motif identification using a dictionary model,” *IEEE Proceedings*, vol. 90, pp. 1803–1810, 2002.

[26] Sabatti, C., L. Rohlin, K. Lange, and J. Liao (2004) “Vocabulon: a dictionary model approach for reconstruction and localization of transcription factor binding sites,” *Bioinformatics*, to appear.

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

[28] West, M. (2003) Bayesian factor regression models in the ”Large p, Small n” paradigm, Bayesian Statistics, 7:723–732.

[29] Yang, Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, T. P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research, 2002, Vol. 30, No. 4 e15.

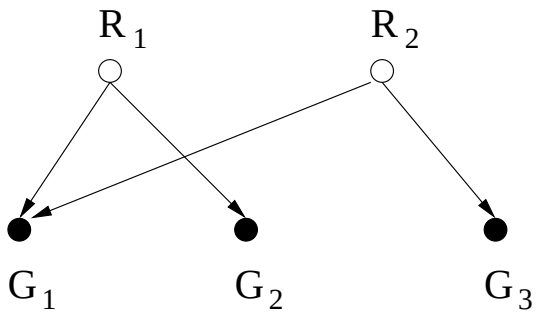

Figure 1: A network for transcription regulations. Parent nodes, indicated as empty circles, represent regulatory proteins, or transcription factors. Descendents, filled circle, targeted genes.

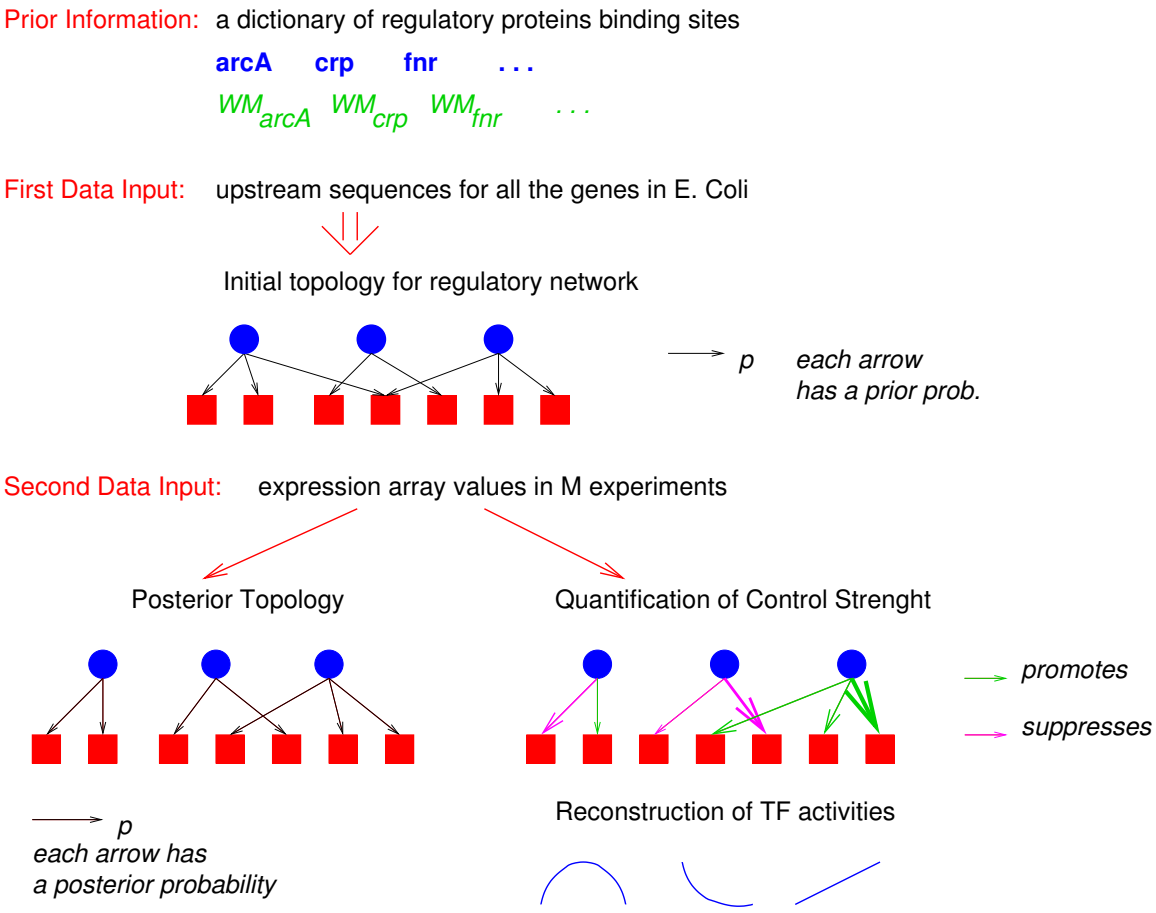

Figure 2: Schematic representation of our algorithm. Initially a known dictionary of transcription factors is used to evaluate all the regulatory regions in E. Coli to identify potential binding sites. This information is combined with the available litterature knowledge to define a prior on the regulatory network. The analysis of expression array data with the model described in this paper leads to a posterior probability on the topology of the network, the quantification of control strength, and the reconstruction of the activation profiles of the transcription factors.

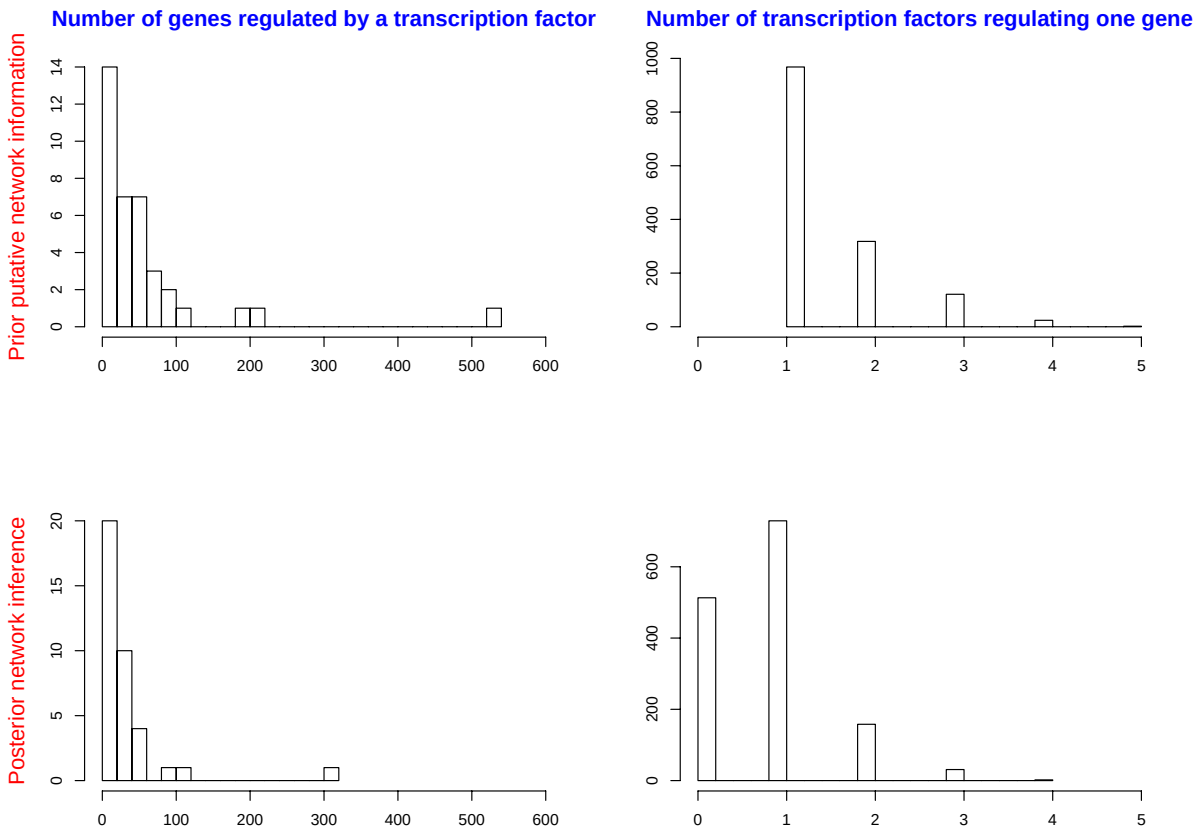

Figure 3: Comparison of the sparsity of prior and posterior network. The top portion of the graph refers to the prior distribution on Z and the lower portion to the posterior. In both cases, we consider as "present" an edge that has probability greater than or equal to 0.5. On the left hand side, we present the distribution of the number of genes regulated by each transcription factor; on the right hand side, the distribution of the number of transcription factors regulating each gene.

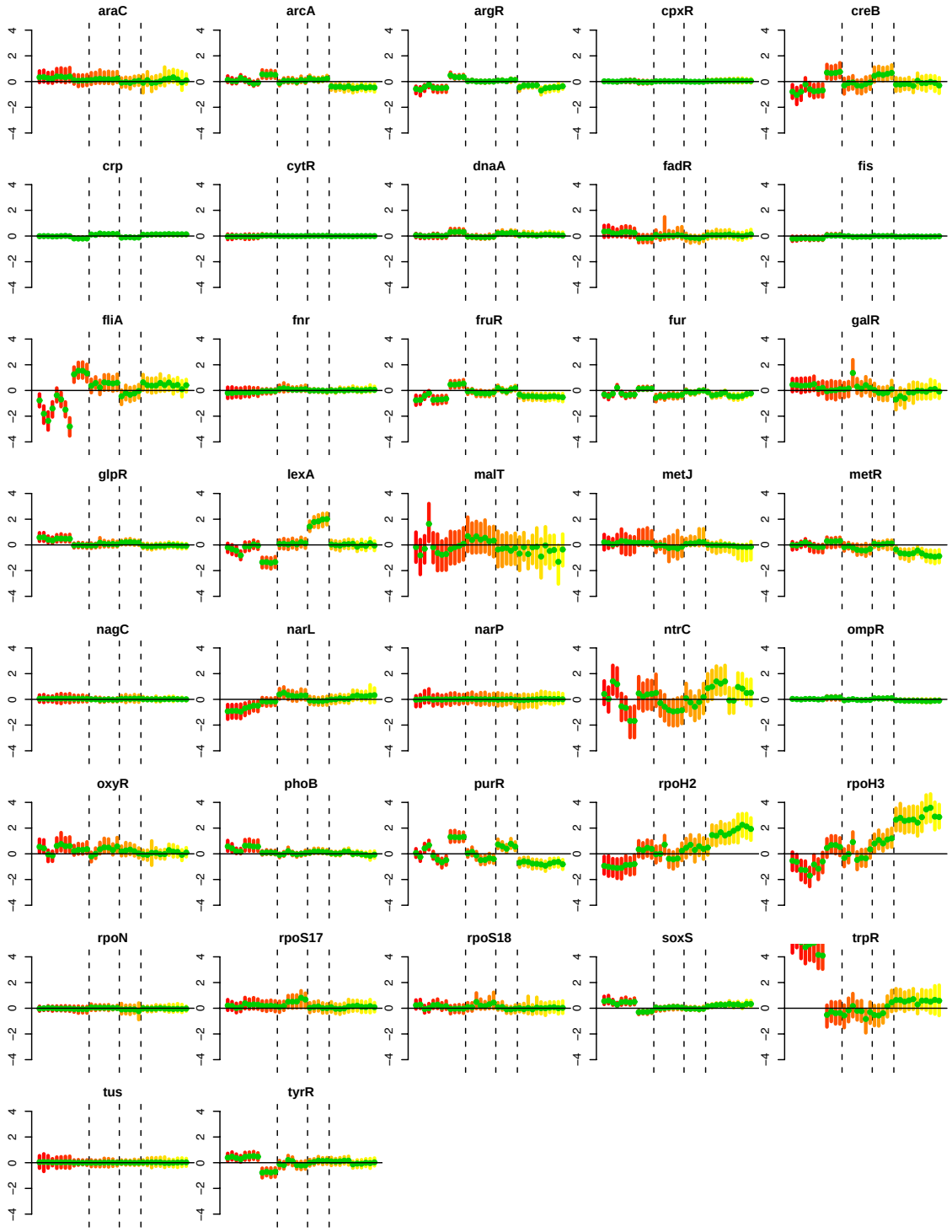

Figure 4: Representation of regulons activities for the 37 transcription factors in the study. Green dots indicate the expected value and the vertical bars span the regions that receive 0.99 probability according to the posterior distribution. Vertical dotted lines are used to separate the four groups of experiments for ease of reading. At the expense of an optimal scale for the visualization of the profile of each transcription factor, we have used the same scale in each of the graphs to aid comparison.

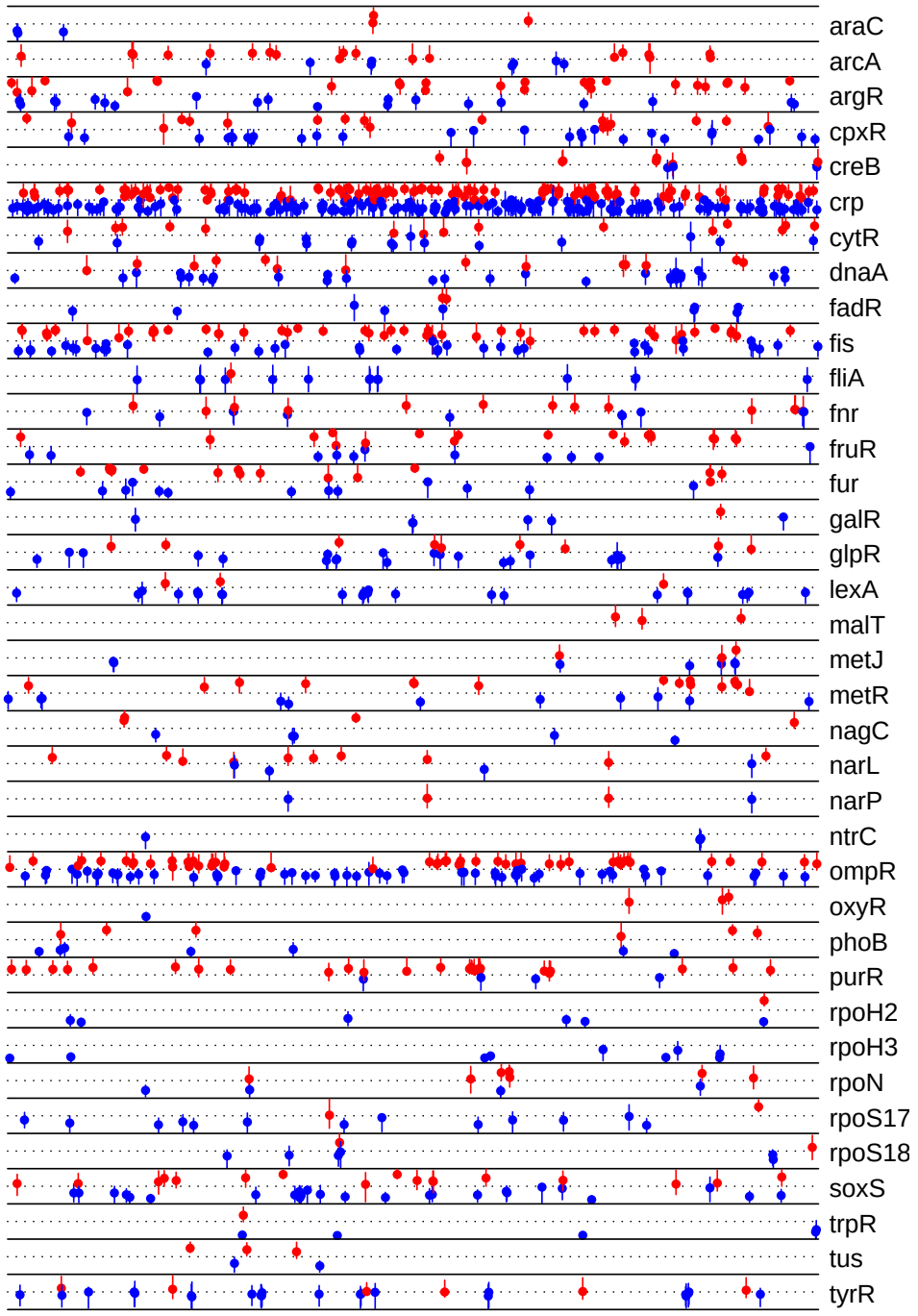

Figure 5: Summary representation of the average impact across experiments of the transcription factors on the genes they appear to regulate (posterior probability of $z_{ij} = 1$ larger than 0.8). Each transcription factor is described in one row display, and each gene correspond to one position on the x axis. Dotted lines indicate the position of zero in the displays. Red and blue are used to indicate effects of opposite signs within the group of genes regulated by the same TF. Confidence intervals (99%) are represented by vertical bars.

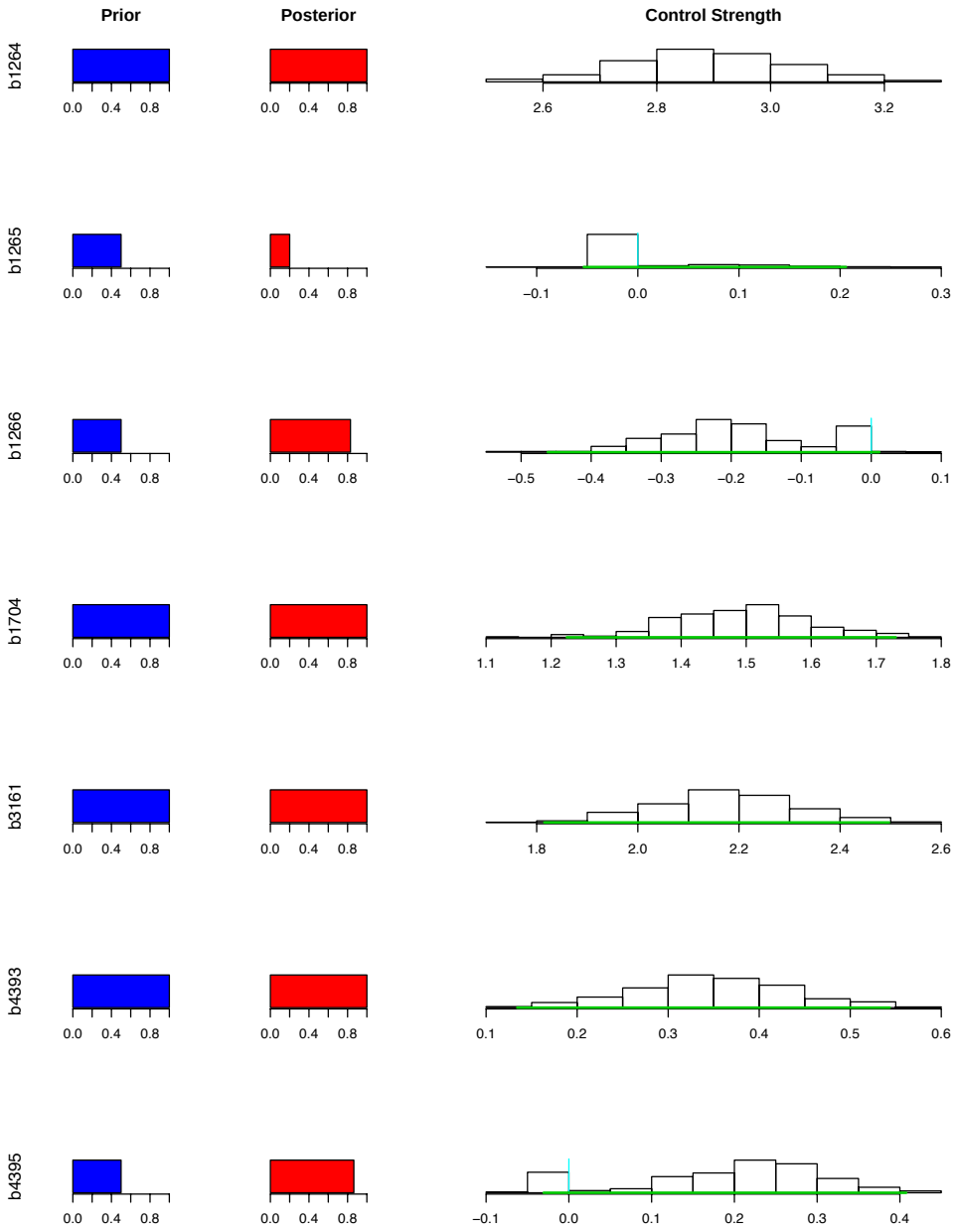

Figure 6: The trpR regulon: connectivity and control strength information. Each row corresponds to one gene that can be potentially regulated by trpR. Genes are indicated by their "b-numbers." The first column represents the initial probability with which trpR is thought to regulate the target genes. 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.

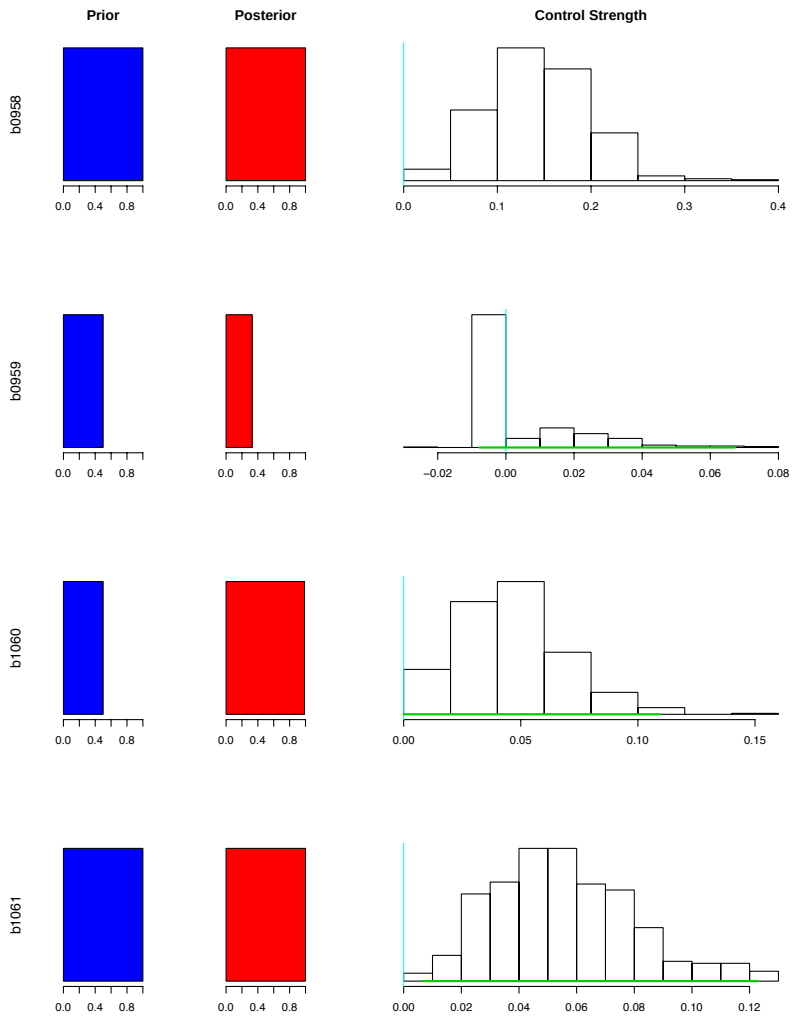

Figure 7: Connectivity and control strength information for part of the lexA regulon. Each row corresponds to one gene that can be potentially regulated by lexA. Genes are indicated by their "b-numbers." The first column represents the initial probability with which lexA is thought to regulate the target 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.