BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–8

Query-driven biclustering of microarray data by Gibbs sampling

Q. Sheng [a*], K. Lemmens [a], K. Marchal [ab], B. De Moor [a], and Y. Moreau [a]

[a]Department of Electrical Engineering, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium
[b]Center of Microbial and Plant Genetics, Katholieke Universiteit Leuven, Kasteelpark Arenberg 20, 3001 Leuven-Heverlee, Belgium

ABSTRACT

Motivation: Existing (bi)clustering methods for microarray data analysis usually reveal global patterns in the data, which often do not answer the questions that are of specific interest to the biologist. Bayesian probabilistic models show promise for pattern discovery directed by a soft query, thanks to their ability to incorporate prior knowledge.

Results: In this paper, we describe a method based on Bayesian probabilistic models to address the biclustering problem when biologists query the data with a set of seed genes that they believe to have a common function. The problem then is to recruit other genes that might have the same function as the seed genes, and in the meantime identify the experimental conditions where this function is active, by finding genes that have similar expression profiles as the seed genes under a subset of experimental conditions. Gibbs sampling is used for the estimation of the model. We demonstrate the efficiency of our method based on a combined data set on *Saccharomyces cerevisiae*.

Availability and supplementary information:
http://www.esat.kuleuven.be/∼qsheng/query_driven.html.

Contact: qizheng.sheng@esat.kuleuven.be

Keywords: Microarray data analysis, Bayesian model, biclustering, clustering, Gibbs sampling

1 INTRODUCTION

Probabilistic models have become a popular choice for modeling microarray data because they handle the high level of noise of microarrays in a principled way. However, the probability distribution provided by these models usually contains many modes, because of the complexity of the underlying biological process. In clustering, these modes correspond to the different clusters that can be identified in the data. The largest clusters (which are easiest to identify) are often not the most interesting to the biologist because they correspond to well-known generic biological functions—where few novel findings are to be expected. This lack of sharpness of clustering algorithms has kept them into a vague exploratory role; because for biologists, one of the main questions is always "what are the genes that are related to a *particular* function (or in a *specific* pathway) of interest to me?"

Bayesian probabilistic models have shown promise in providing answers to this type of question, by transforming the existing knowledge of biologists into the prior probabilities that are incorporated into the model (Segal *et al*., 2001, 2003). Because in Bayesian models the likelihood of the data is multiplied by the prior to deliver the posterior probabilistic model, a proper prior can substantially raise the mode that is most relevant to the biological question in the posterior model.

Coming back to the clustering problem for microarray data, and supposing that the biologists have at hand a specific set of genes (called the "seed genes" hereafter), which they know to be related to some common biological function, the question for their query to the microarray data is "which other genes in this data set share similar expression profiles as the seed genes and thus might be involved in the same function?"

We generalize this question further by considering which data set could be considered to answer such a question. Until recently, clustering would be performed on the array data from a single study addressing a limited biological situation. Yet, in the last few years, large data sets (called microarray compendia) consisting of multiple biological conditions or data from multiple studies have demonstrated their effectiveness as the basis of guilt-by-association studies. So in a complex compendium, where it may not be clear which microarray conditions are truly most relevant to the biological question at hand, the question becomes "which genes are functionally related to the seed genes, and in the meantime, in which experimental conditions is this biological function involved?" Otherwise stated, given the seed genes, we want to recruit genes (presented in the microarray data set) that share similar expression profiles under a subset of conditions. In addition, the few seed genes whose profiles are not compatible with the discovered pattern should be rejected if present. This is what we call the "query-driven biclustering" problem. (See Figure 1 for an illustration.)

Another similar problem also exists for the other orientation of microarray data. For example, given a set of patients who share a certain pathological similarity, the question is to recruit other patients of the same type, and in the meantime, to identify the genes that provide a fingerprint that characterize those patients. Hereafter, we will refer to the former problem as the query-driven biclustering of genes, and to the latter one as the query-driven biclustering of experiments.

Biclustering techniques for microarray data, pioneered by Cheng and Church (2000), are receiving increasing attention in bioinformatics. Unlike conventional clustering algorithms, the biclustering algorithms (see Madeira and Oliveira (2004) for a survey) aim at finding genes that show consistent behavior only under a *subset* of experiments. The discovery of such relationship between the

*To whom correspondence should be addressed.

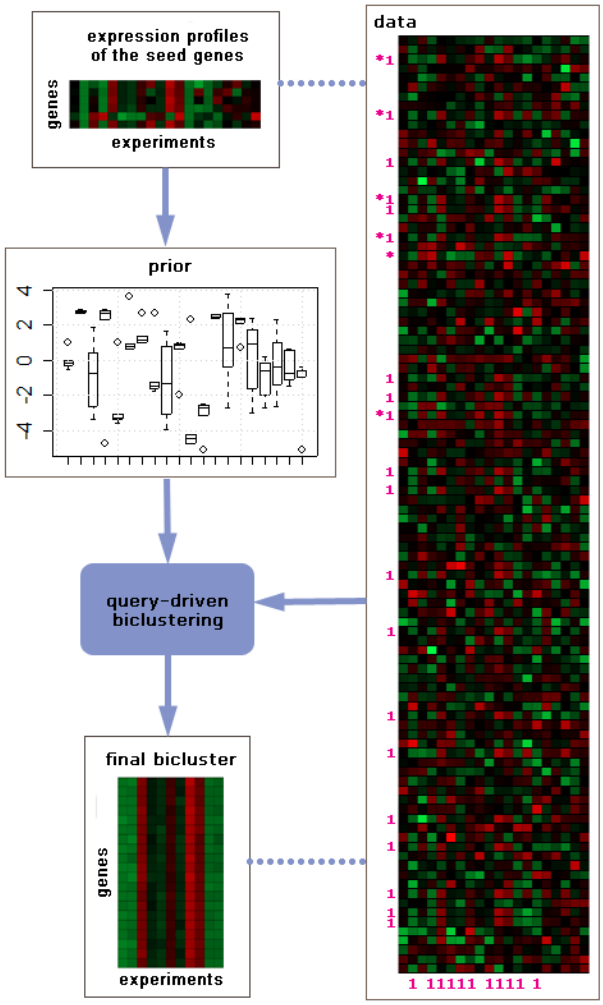

Fig. 1. Conceptual scheme of query-driven biclustering. The rows labelled with "*" represent the expression profiles of the seed genes, and the rows and columns labelled with "1" represent genes and experiments that belong to the discovered bicluster.

genes and the conditions provides crucial information for unveiling genetic pathways. However, most of the existing biclustering algorithms focus on revealing the global pattern of the data instead of being motivated by a specific biological query. Yet, one example of addressing problems similar to the query-driven biclustering of genes is the signature algorithm (Ihmels *et al*., 2002), which is based on the correlation measure between a gene and the average profile of the seed genes under experimental conditions where the average expression value (of the seed genes) is above a threshold.

In our previous paper (Sheng *et al*., 2003), we described a Bayesian hierarchical model for discretized data using a multinomial likelihood and a Dirichlet prior for tackling the biclustering problem of patients. The same model can be used for the query-driven biclustering of patients by imposing a tailored Dirichlet prior. (Because of the lack of space, see the supplementary material for an example.) In this paper, we describe a Bayesian hierarchical model based on Gaussian likelihood and Gaussian-Wishart prior to address the query-driven biclustering of genes on continuous microarray data. We use a combined data set from Gasch *et al*. (2000), Spellman *et al*. (1998) and Cho *et al*. (1998) to illustrate the efficiency of our method.

2 MODELS AND APPROACHES

Let us assume that a microarray data set contains only one bicluster, hence the rest of the data is considered as background noise. To put the problem in a probabilistic manner, we use a binary random variable (called a label) to describe whether a gene or an experiment belongs to the bicluster (i.e., a "1" indicates that the gene or experiment is in the bicluster and a "0" indicates otherwise). Considering the labels as the hidden data of the problem (and the microarray data as the observed data), the biclustering problem becomes two-fold: (1) given a fixed assignment of the labels, we want to find the most suitable probabilistic model for the data and (2) given the models of the bicluster and the background, assess the value of the labels. Both expectation-maximization (EM) (Barash and Friedman, 2002) and Gibbs sampling are algorithms to solve such data augmentation problems.

Gibbs sampling is a Markov chain Monte Carlo (MCMC) method for joint distribution estimation. The joint distribution concerns both the hidden data and the parameters of the model, both of which are considered as random variables for the Gibbs sampling procedure. Given the full conditional distribution for each random variable, the Gibbs sampling procedure iteratively draws samples from the conditional distributions. The samples drawn by this procedure are guaranteed to converge to the joint distribution (Casella and George, 1992). Once the converged samples are collected, the *maximum a posteriori* (MAP) estimation of a random variable can be obtained by performing Monte Carlo integration,

$$\mathrm{E}[x_i|\mathcal{D}] = \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}[x_i \mid x_1^{(t)} \ldots x_{i-1}^{(t)}, x_{i+1}^{(t)} \ldots x_m^{(t)}, \mathcal{D}], \quad (1)$$

where the x_i's, for all $i \in \{1, \ldots, m\}$, (m is the total number of random variables) are the random variables under concern, $\mathcal{D}$ denotes the data, t indexes the iterations (performed after the procedure converges), and T is the total number of samples in the sample pool.

Applying Gibbs sampling to the biclustering problem of microarray data (Sheng *et al*., 2003) directly to the previous model often results in revealing a large global bicluster embedded in the data – unless a strong prior is imposed on the model to zoom in on a specific part of the data.

In this section, we discuss the Gibbs sampling strategy for tackling the query-driven biclustering problem of genes in the following five aspects:

- *Hidden data*: the structure and the prior distribution of the labels.
- *Data models*: the hierarchical Bayesian models describing the microarray data in both the bicluster and the background.
- *Full conditional distributions*: the distributions from which samples of the labels and samples of the model parameters are drawn during the Gibbs sampling procedure.
- *Construction of the priors*: the incorporation of information from the seed genes into the Bayesian hierarchical models.

- *Gibbs sampling procedure*: a final overview of the Gibbs sampling scheme.

2.1 Hidden data

From now on, we distinguish between the phrases "experiment" and "condition" by defining an experiment as a column of the microarray data matrix, and a condition as a group of experiments. This distinction is useful, for example, when the microarray data is obtained from time-series experiments. In this case, different columns in a microarray data set may correspond to experiments that are performed under the same condition but at different time points. When performing the biclustering algorithm, we might want to assign experiments from the same condition to one bicluster by using one label to describe the association of these experiments to the bicluster, yet in the meantime, we would use different model parameter settings to describe different experiments.

For this reason, we use one label per condition to describe whether the corresponding group of experiments belongs to the cluster. However, for the genes, we use one label per gene to indicate its relation with the bicluster.

As for the notations, we use capital letters to denote the labels; i.e.,

$$\mathbf{G} = \big\{G_i \in \{0,1\} \,|\, i = 1, \ldots, n\big\} \quad (2)$$

$$\mathbf{C} = \big\{C_j \in \{0,1\} \,|\, j = 1, \ldots, q\big\} \quad (3)$$

respectively for the genes and the conditions, where n is the number of genes in the data set, and q is the number of conditions.

We use up-right bold lowercase letters to denote sets of indices. More specifically, we use

$$\mathbf{g} = \{i \,|\, G_i = 1 \wedge G_i \in \mathbf{G}\} \quad (4)$$

$$\bar{\mathbf{g}} = \{i \,|\, G_i = 0 \wedge G_i \in \mathbf{G}\} \quad (5)$$

to denote respective the indices of genes in the bicluster and the indices of genes in the background. For the column dimension of a microarray data matrix, we only need the indices of the experiments for the evaluation of our model. We use $\mathbf{e}_j$ to specify the indices of the experiments under condition j. To denote the whole set of indices of the experiments whose corresponding conditions are in the bicluster or in the background, we use respectively

$$\mathbf{e} = \big\{k \,|\, k \in \mathbf{e}_j, \forall j = \{1, \ldots, q\} \wedge C_j = 1\big\}, \quad (6)$$

$$\bar{\mathbf{e}} = \big\{k \,|\, k \in \mathbf{c}_j, \forall j = \{1, \ldots, q\} \wedge C_j = 0\big\}. \quad (7)$$

A dot, "$\cdot$" is used to refer to the entire set of indices in one dimension of the microarray data matrix.

In addition, when a subscription $\bar{i}$ is added to $\mathbf{G}$ or $\mathbf{g}$ (or when $\bar{j}$ is added to $\mathbf{C}$) it indicates that the set include all but the i^{th} genes (or all but the j^{th} conditions).

To put the problem in the Bayesian framework means that the hidden data as well as the observed data are modeled by Bayesian models. For the labels, we use

$$\lambda_{\text{g}} \equiv P(G_i = 1), \quad \forall i = 1 \ldots n \qquad \lambda_{\text{g}} \sim \text{Beta}(\xi_0^{\text{g}}, \xi_1^{\text{g}}) \quad (8)$$

$$\lambda_{\text{c}} \equiv P(C_j = 1), \quad \forall j = 1 \ldots q \qquad \lambda_{\text{c}} \sim \text{Beta}(\xi_0^{\text{c}}, \xi_1^{\text{c}}), \quad (9)$$

where λ_{g} and λ_{c} are parameters of the Bernoulli prior that a gene (or a condition) belongs to the bicluster. Both λ_{g} and λ_{c} are further modeled by two corresponding Beta distributions.

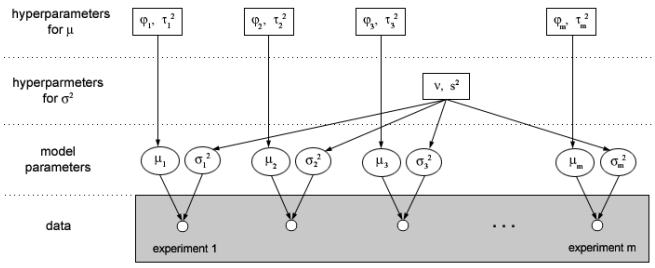

Fig. 2. Data model for the query-driven biclustering of genes. The same model structure is used to describe both the bicluster and the background.

2.2 Data model

Given a fixed division of the microarray data (into the bicluster and the background), we use the same hierarchical Bayesian model structure to describe both the data in the bicluster and the data in the background, which is illustrated in Figure 2. In either of the models, we use normal distributions to describe the expression data. This choice is based on not only the analytical convenience of normal models, but also the previous success in applying normal mixture models to the clustering problems of microarray data (Yeung *et al*., 2001; McLachlan *et al*., 2002). Furthermore, the assumption for fitting a normal distribution to the gene expression measurements in a given situation is considered to be reasonable especially when a proper preprocessing procedure has been applied to the microarray data (Baldi and Long, 2001).

The model structure implies the assumption that measurements under different experiments are conditionally independent of each other given the prior on the model. Gene expression measurements under each experiment are modeled by a single normal distribution with mean μ_k and variance σ_k^2 ($k \in \mathbf{e}$ when applied to the bicluster model, and $k = 1 \ldots m$ when applied to the background model). To distinguish the bicluster model from the background model, we use $\boldsymbol{\mu}^{\text{bcl}}$ and $(\boldsymbol{\sigma}^2)^{\text{bcl}}$ to denote the parameters of the bicluster model, and $\boldsymbol{\mu}^{\text{bgd}}$ and $(\boldsymbol{\sigma}^2)^{\text{bgd}}$ for those of the background.

For either of the models, the conditional independency means that the correlation between the measurements under different experiments in the bicluster (or in the background) is explained by the prior on $(\boldsymbol{\sigma}^2)^{\text{bcl}}$ (or $(\boldsymbol{\sigma}^2)^{\text{bgd}}$). We use conjugate priors for $\boldsymbol{\sigma}^2$ (i.e., the same scaled inverse-χ^2 distribution with degree of freedom ν and scale s^2 for all the σ_k^2 in $\boldsymbol{\sigma}^2$). Conjugate priors are also used for $\boldsymbol{\mu}$. In this case, to provide flexibility to the model, each μ_k has a different prior distribution, which is also a normal distribution with mean φ_k and variance τ_k^2.

Given the association of the genes and the conditions to the bicluster and the background, as well as the fixed parameters for both the bicluster model and the background model, a gene expression profile is assumed to be i.i.d. drawn from a combination of the bicluster model and the background model when the gene belongs to the bicluster, or from a pure background model when the gene belongs to the background.

The parameters of the priors on $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ are all considered as hyperparameters, which means that they are only used as input of the algorithm. The parameterization procedure are carried out for the inference of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ only.

2.3 Full conditional distributions

The target joint distribution of the query-driven biclustering problem therefore is $P(\mathbf{G}, \mathbf{E}, \lambda_{\mathrm{g}}, \lambda_{\mathrm{c}}, \mathcal{M} \mid \mathcal{D})$, where $\mathcal{D}$ stands for the microarray data (with n genes, and m experiments grouped into q conditions), and $\mathcal{M}$ represents the whole set of model parameters considered in the Bayesian hierarchical models for the both the bicluster and the background; i.e., $\{\boldsymbol{\mu}^{\mathrm{bcl}}, (\boldsymbol{\sigma}^2)^{\mathrm{bcl}}, \boldsymbol{\mu}^{\mathrm{bgd}}, (\boldsymbol{\sigma}^2)^{\mathrm{bgd}}\}$. This means that in the Gibbs sampling procedure, we need to iteratively sample from the full conditional distribution of each of the involved random variables. However, as shown in Sheng *et al*. (2003), we can simplify the sampling procedure by integrating out λ_{g} and λ_{c}. Therefore, the joint distribution of interest becomes $P(\mathbf{G}, \mathbf{E}, \mathcal{M} \mid \mathcal{D})$.

Let us begin with the conditional distributions of the labels. Because the labels are Bernoulli variables by definition, their full conditional distributions are in the form of Bernoulli distributions. Instead of evaluating the Bernoulli parameter λ, we use the odds γ between λ and $1-\lambda$ to characterize the full conditional distribution of a label,

$$\gamma = \frac{\lambda}{1-\lambda}. \tag{10}$$

For a gene label,

$$\begin{aligned}\gamma_i^{\mathrm{g}} &= \frac{P(G_i = 1 \mid G_{\bar{i}}, \mathbf{C}, \mathcal{D}, \mathcal{M})}{P(G_i = 0 \mid G_{\bar{i}}, \mathbf{C}, \mathcal{D}, \mathcal{M})} \\ &= \frac{P(\mathcal{D}, G_i = 1, \mathbf{G}_{\bar{i}}, \mathbf{C} \mid \mathcal{M})}{P(\mathcal{D}, G_i = 0, \mathbf{G}_{\bar{i}}, \mathbf{C} \mid \mathcal{M})}, \quad i \in \{1, \ldots, n\},\end{aligned} \tag{11}$$

Similarly, for a condition label, we have

$$\begin{aligned}\gamma_j^{\mathrm{c}} &= \frac{P(C_j = 1 \mid \mathbf{C}_{\bar{j}}, \mathbf{G}, \mathcal{D}, \mathcal{M})}{P(C_j = 0 \mid \mathbf{C}_{\bar{j}}, \mathbf{G}, \mathcal{D}, \mathcal{M})} \\ &= \frac{P(\mathcal{D}, C_j = 1, \mathbf{C}_{\bar{j}}, \mathbf{G} \mid \mathcal{M})}{P(\mathcal{D}, C_j = 0, \mathbf{C}_{\bar{j}}, \mathbf{G} \mid \mathcal{M})}, \quad j \in \{1, \ldots, q\},\end{aligned} \tag{12}$$

Equations 11 and 12 imply that with a larger likelihood ratio of the label, the gene (or the experiment) has a larger probability to be in the bicluster, and that a smaller likelihood ratio suggests otherwise.

Note that for the evaluation of Equation 11 and 12

$$\begin{aligned}&P(\mathcal{D}, \mathbf{G}, \mathbf{C} \mid \mathcal{M}) \\ =&P(D \mid \mathbf{G}, \mathbf{C}, \mathcal{M}) \cdot \int_{\lambda_{\mathrm{g}}} \int_{\lambda_{\mathrm{c}}} P(\mathbf{G} \mid \lambda_{\mathrm{g}}) \cdot P(\mathbf{C} \mid \lambda_{\mathrm{c}}) \, \mathrm{d}\lambda_{\mathrm{g}} \, \mathrm{d}\lambda_{\mathrm{c}}.\end{aligned} \tag{13}$$

Because of the i.i.d. distribution of the gene expression profiles and the conditional independence of the normal distributions, the likelihood ratio γ_i^{g} can be evaluated only on the data of the concerned gene and under the conditions that are currently assigned to the bicluster,

$$\begin{aligned}\gamma_i^{\mathrm{g}} &= \frac{P\big(\mathcal{D}[i, \mathbf{e}] \mid \boldsymbol{\mu}^{\mathrm{bcl}}, (\boldsymbol{\sigma}^2)^{\mathrm{bcl}}\big)}{P\big(\mathcal{D}[i, \mathbf{e}] \mid \boldsymbol{\mu}^{\mathrm{bgd}}, (\boldsymbol{\sigma}^2)^{\mathrm{bgd}}\big)} \cdot \frac{|\mathbf{g}_{\bar{i}}| + \xi_1^{\mathrm{g}}}{n - 1 - |\mathbf{g}_{\bar{i}}| + \xi_0^{\mathrm{g}}}, \\ & i \in \{1, \ldots, n\},\end{aligned} \tag{14}$$

where we use a pair $\mathbf{u}$ and $\mathbf{v}$ in $\mathcal{D}[\mathbf{u}, \mathbf{v}]$ to indicate the part of the data under concern, with $\mathbf{u}$ providing the indices of genes and $\mathbf{v}$ providing the indices of experiments. The second term in Equation 14 results from the integration of λ_{g}. Similarly, for the conditions, we have

$$\begin{aligned}\gamma_j^{\mathrm{c}} &= \frac{P\big(\mathcal{D}[\mathbf{g}, \mathbf{e}_j] \mid \boldsymbol{\mu}^{\mathrm{bcl}}, (\boldsymbol{\sigma}^2)^{\mathrm{bcl}}\big)}{P\big(\mathcal{D}[\mathbf{g}, \mathbf{e}_j] \mid \boldsymbol{\mu}^{\mathrm{bgd}}, (\boldsymbol{\sigma}^2)^{\mathrm{bgd}}\big)} \cdot \frac{w_{\bar{j}} + \xi_1^{\mathrm{c}}}{q - 1 - w_{\bar{j}} + \xi_0^{\mathrm{c}}}, \\ & j \in \{1, \ldots, q\},\end{aligned} \tag{15}$$

where $w_{\bar{j}}$ denotes the number of conditions in the current bicluster. Note that by using likelihood ratios, the missing values in the microarray data can be neglected from the evaluation of the conditional distributions, which is equivalent to assuming that these data points have the same possibility be to generated by the bicluster model as by the background model.

Using conjugate priors for the model parameters means that their conditional distributions are in the same form as the prior. In either the bicluster model or the background model, the conditional distribution for μ_k remains a normal distribution,

$$\begin{aligned}&p(\mu_k \mid \sigma_k^2, \mathbf{G}, \mathbf{C}, \mathcal{D}) = N(\hat{\mu}_k, \hat{\sigma}_k^2) \\ &\hat{\mu}_k = \frac{\frac{\varphi_k}{\tau_k^2} + \frac{\bar{\mu}_k}{\frac{\sigma_k^2}{a_k}}}{\frac{1}{\tau_k^2} + \frac{1}{\frac{\sigma_k^2}{a_j}}} \quad \text{and} \quad \hat{\sigma}_k^2 = \frac{1}{\frac{1}{\tau_k^2} + \frac{1}{\frac{\sigma_k^2}{a_k}}}.\end{aligned} \tag{16}$$

The posterior distributions for σ_k^2 is a scaled inverse-χ^2 distribution,

$$\begin{aligned}&p(\sigma_k^2 \mid \mu_k, \mathbf{G}, \mathbf{C}, \mathcal{D}) = \text{Inverse-}\chi^2(\hat{\nu}, \hat{s}^2) \\ &\hat{\nu} = \nu + a_k \quad \text{and} \quad \hat{\sigma}^2 = \frac{(a_k - 1) \cdot \bar{s}^2 + \nu \cdot s^2}{\hat{\nu}}.\end{aligned} \tag{17}$$

In Equation 16 and 17, $\bar{\mu}_k$ and $\bar{s}_k^2$ denote respectively the sample mean and sample variance of the relevant data; for $\boldsymbol{\mu}^{\mathrm{bcl}}$ and $(\boldsymbol{\sigma}^2)^{\mathrm{bcl}}$

$$k \in \mathbf{e} \quad \text{and} \quad a_k = |\mathbf{g}|,$$

and for $\boldsymbol{\mu}^{\mathrm{bgd}}$ and $(\boldsymbol{\sigma}^2)^{\mathrm{bgd}}$

$$\begin{aligned}&k = 1 \ldots q, \\ &a_k = \begin{cases} |\bar{\mathbf{g}}| & k \in \mathbf{e} \\ n & k \in \bar{\mathbf{e}} \end{cases}.\end{aligned}$$

Note that the missing values in the microarray data can also be left out of the evaluation of posterior distributions for the model parameters.

2.4 Construction of the priors

To impose our requirement that the mean of the genes under each experiment in the bicluster should strictly follow that of the mean of seed genes, we set

$$\boldsymbol{\varphi}^{\mathrm{bcl}} = \boldsymbol{\varphi}'[\mathbf{e}], \tag{18}$$

where $\boldsymbol{\varphi}'$ is calculated as the mean of the seed genes under all the experiments in the data set, and we use a very small value for $\boldsymbol{\tau}^{\mathrm{bcl}}$, for example,

$$\tau_k^{\mathrm{bcl}} = 10^{-4}, \, k \in \mathbf{e}. \tag{19}$$

By setting

$$(s^2)^{\mathrm{bcl}} = \frac{1}{\nu^{\mathrm{bcl}}} \tag{20}$$

for the prior on $(\boldsymbol{\sigma}^2)^{\mathrm{bcl}}$, the scaled inverse-$\chi^2$distribution becomes an inverse-χ^2 distribution, which means that no prior knowledge

on the exact value of the posterior variance is imposed, and that the posterior parameters for $(\boldsymbol{\sigma}^2)^{\text{bcl}}$ are of smaller values for those experiments under which the selected genes have a smaller sample variance. Raising ν^{bcl} implies stronger belief that the posterior variance is close to the sample variance of the selected genes, the effect of which is equivalent to increasing the number of genes in the bicluster without changing the sample variance.

For the prior on $\boldsymbol{\mu}^{\text{bgd}}$, we set $\boldsymbol{\varphi}_k^{\text{bgd}}$ to the mean of the expression levels of all the genes under experiment k. If the data under each experiment is rescaled to have unit variance before the query-driven biclustering analysis, we set

$$\tau_k^{\text{bgd}} = 1, \quad k = 1 \ldots m. \tag{21}$$

Otherwise, a weak prior can be used by setting τ_k^{bgd} to a large value, such as

$$\tau_k^{\text{bgd}} = 10^4, \quad k = 1 \ldots m. \tag{22}$$

For the priors on $(\boldsymbol{\sigma}^2)^{\text{bgd}}$, typically, we set

$$\nu^{\text{bgd}} = 0.01n \tag{23}$$

$$(s^2)^{\text{bgd}} = \frac{1}{\nu^{\text{bgd}}}. \tag{24}$$

In addition, weak priors are also used for the labels, because we have little knowledge beforehand about how many genes and conditions the bicluster would contain. We typically set

$$\xi_0^{\text{g}} = \xi_1^{\text{g}} = 0.5 \tag{25}$$

$$\xi_0^{\text{c}} = \xi_1^{\text{c}} = 0.5. \tag{26}$$

In this way, ν^{bcl} is the only hyperparameter that is open to the user for controlling the stringency of the bicluster.

2.5 Gibbs sampling procedure

We initialize the labels of the seed genes to 1, and the rest of the gene labels to 0. On the other hand, we initialize the condition labels randomly either to 1 or to 0. The model parameters for both the bicluster and the background are initialized in accordance with their priors (i.e., initializing $\boldsymbol{\mu}$ to $\boldsymbol{\varphi}$, and $\boldsymbol{\sigma}^2$ to s^2).

During the Gibbs sampling procedure, the labels and model parameters are sampled one at a time from their full conditional distributions (see Section 2.3) iteratively. After the sample statistics converge to the joint distribution, some additional iterations are performed (still by the Gibbs sampling procedure) for which the samples of the labels are collected. The final Bernoulli parameters of the labels are evaluated as described in Equation 1.

Although the model parameters are sampled during the Gibbs sampling procedure, in order to reduce the memory complexity of the algorithm, the samples of these parameters are not collected. Instead, sample statistics will be used to inspect the model of both the bicluster and the background, after the position of the bicluster is determined.

3 DATA AND RESULTS

The query-driven biclustering algorithm is applied to the combined data set on *Saccharomyces cerevisiae* from Gasch *et al*. (2000) (with stress-response experiments), Spellman *et al*. (1998) and Cho *et al*. (1998) (both with cell-cycle-related experiments). Each of the original data set was centered and rescaled so that measurements under every microarray experiment have mean of 0 and standard deviation of 1. Then, the gene profiles in each data set were centered and rescaled in the same way. The resulting data sets were then put alongside of each other.

De Bie *et al*. (2005) describes a method to combine three independent data sources, namely genome-wide location data (ChIP chip data), motif information as obtained by phylogenetic shadowing, and gene expression profiles, for the construction of a biologically meaningful set of seed genes for a certain transcription module. Seed genes are identified from the input information as those that share the same combination of regulators and motifs, and whose expression profiles have a large correlation. We use three sets of seed genes found by their method to examine the effectiveness of our method, especially to test the influence of ν^{bcl} on the final bicluster. Two sets of seed genes (referred to as Seed 1 and Seed 2 hereafter) are composed of cell cycle related genes, for which we expect that the algorithm would identify all the experimental conditions under the data from Spellman *et al*. (1998) and Cho *et al*. (1998), and recruit additional genes related to cell cycle regulation. The other set of seed genes (i.e., Seed 3) is involved in ribosome biogenesis, a more general function for which we expect the algorithm to find a bicluster consisting of most of the experimental conditions in the data set.

For each set of seed genes, we report three different values of ν^{bcl} in Table 1 and Table 2. Each time, we ran the biclustering algorithm for 1000 iterations, and the number of burn-in iterations (i.e., iterations before convergence is reached and are not taken into account for the final evaluation) was determined as described in Sheng *et al*. (2003). A gene or a condition was selected (to be in the bicluster) if in 95% of the collected samples (i.e., iterations), the gene or the condition had a probability of more than 0.9 to be in the bicluster. Finally, we validated the bicluster by calculating the functional enrichment of the bicluster using a hypergeometric distribution (Tavazoie *et al.*, 1999), where the functional categories of the genes are obtained from MIPS (Mewes *et al*., 2004). In Table 1 and 2 we only report the functional categories whose p-values are lower than 0.001 (as well as those p-values).

4 DISCUSSION

In general, the increment of ν^{bcl} results in more stringent biclusters, in the sense that both the number of genes and the number of conditions drop with a larger ν^{bcl}, and that the selected genes cover fewer functional categories. From the mathematical point of view, a relatively large ν^{bcl} demands that the conditions to be selected are those under which the selected genes have both a small variation and a similar profile to that of the seed genes. However, the requirement on the variation becomes more stringent when fewer genes are included in the bicluster. That is why sometimes a condition is not included in the bicluster even if it seems to meet both of the above requirements when using a relatively large ν^{bcl}. On the contrary, when a smaller ν^{bcl} is used, an expression profile similar to that of the seed genes together with a variance slightly smaller than that of the background will be sufficient for a condition to be included in the bicluster. Figure 3A and 3B illustrate the influence of a small and a large ν^{bcl} using Seed 1. When ν^{bcl} gets extremely small, the algorithm will find the global bicluster (in this case, more than two thousand genes under all the experimental conditions). When ν^{bcl} is extremely large, the Gibbs sampling procedure is not able to

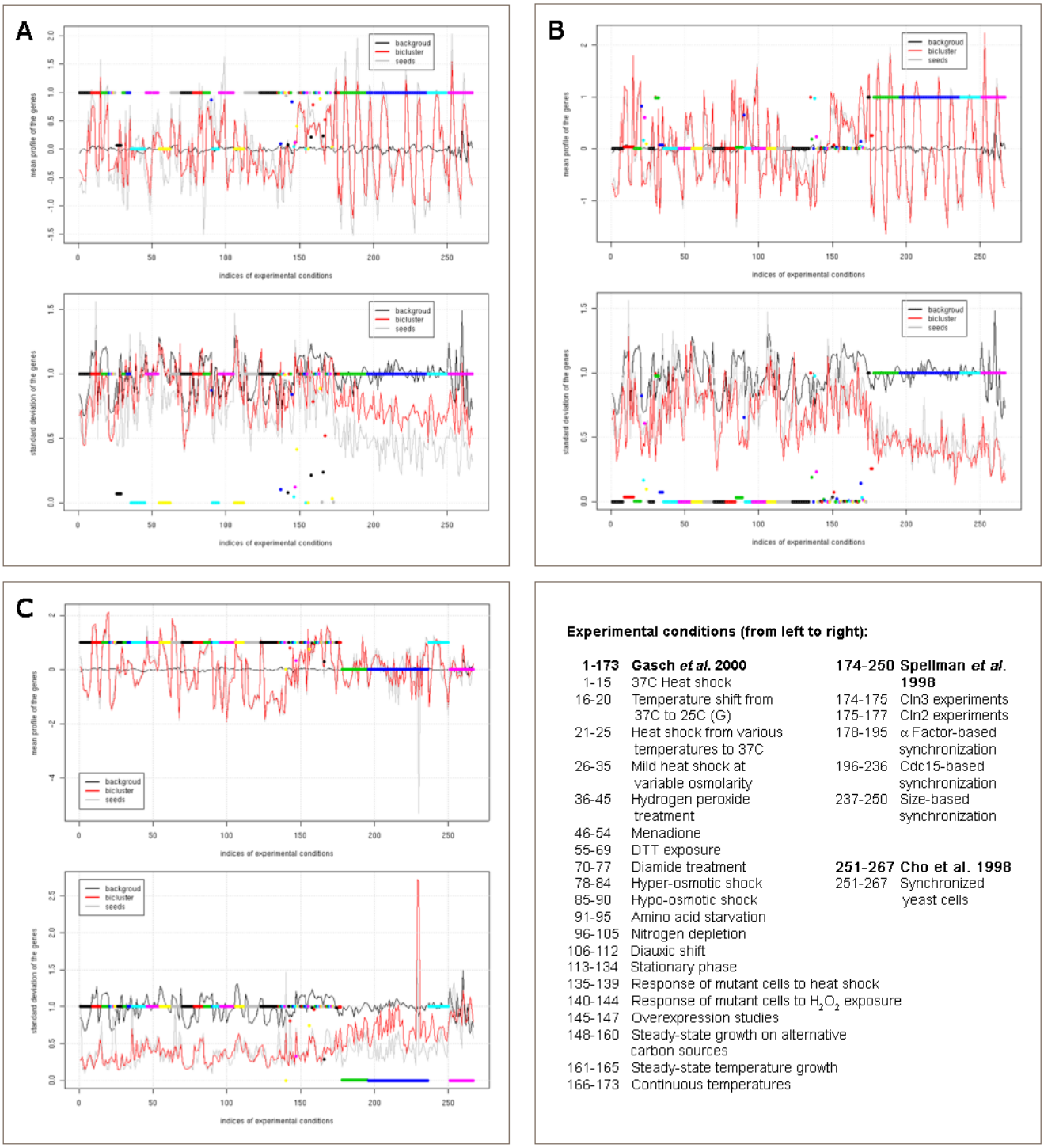

Fig. 3. Plots of the mean expression profiles (upper graphs) and the standard deviations (lower graphs) of the genes (in the bicluster, the seeds and the background) over all the experimental conditions. The color dots in the graphs show the posterior probability for the corresponding condition to be in the bicluster. Experiments are grouped into one condition if they belong to the same time-series experiment, and 8 colors are recycled to represent the 70 groups of experiments (i.e. 70 conditions). (A) Results for Seed1, $\nu^{\text{bcl}} = 20$. (B) Results for Seed 1, $\nu^{\text{bcl}} = 50$. (C) Results for Seed 3, $\nu^{\text{bcl}} = 150$.

Table 1. Influence of ν^{bcl} on the biclustering results.

Nr.	Seed genes	ν^{bcl}	Selected genes*	Excluded seed genes	Selected conds.
1	YCR065W **11.02.03.04** YDR097C **10.01.05** YDL003W **10.03.01** **10.03.04.03** YGR109C **10.01.03** **10.03.01** **10.03.02** YGR221C **40.01** **42.04** **43.01.03.05** YGL038C **01.05.01** **14.07**	20	*187 genes covering 89 functional categories:* **10** Cell cycle and DNA processing (74: 8.47e-13) **10.01** DNA processing (40: 1.13e-12) **10.01.03** DNA synthesis and replication (25: 1.51e-12) **10.01.05** DNA recombination and DNA repair (22: 5.56e-8) **10.01.05.01** DNA repair (13: 2.67e-6) **10.03** cell cycle (46: 1.07e-10) **10.03.01** mitotic cell cycle and cell cycle control (39: 8.47e-13) **10.03.01.03** cell cycle checkpoints (7: 1.15e-5) **10.03.04** nuclear and chromosomal cycle (7: 5.53e-4) **42.04** cytoskeleton (12: 5.54e-4) **43** Cell type differentiation (24: 1.49e-4) **43.01** fungal/microorganismic cell type differentiation (24: 1.49e-4) **43.01.03** fungal and other eukaryotic cell type differentiation (24: 1.49e-4) **43.01.03.05** budding, cell polarity and filament formation (19: 8.51e-5)	*none*	*47 conditions covering:* a large range of conditions in the *Gasch* data set all the conditions in the *Spellman* data set the condition of *Cho* data set
	YER095W **10.01.05** **10.01.05.01** **10.03.02** **34.11.03.07** YLR103C **10.01.03** **10.03.01** YJL074C **10.03.01** **10.03.04** YJL187C **10.03.01.03** **40.01** YPL267W **42.04** **43.01.03.05** **99**	40	*83 genes in 60 functional categories:* **10** Cell cycle and DNA processing (40: 6.87e-12) **10.01** DNA processing (22: 3.42e-10) **10.01.03** DNA synthesis and replication (15: 7.31e-12) **10.01.05** DNA recombination and DNA repair (11: 5.03e-5) **10.01.05.01** DNA repair (6: 1.36e-3) **10.03** cell cycle (25: 3.76e-8) **10.03.01** mitotic cell cycle and cell cycle control (22: 1.04e-9) **10.03.04** nuclear and chromosomal cycle (5: 4.57e-4) **43** Cell type differentiation (13: 8.81e-4) **43.01** fungal/microorganismic cell type differentiation (13: 8.81e-4) **43.01.03** fungal and other eukaryotic cell type differentiation (13: 8.81e-4) **43.01.03.05** budding, cell polarity and filament formation (11: 3.19e-4)	YER095W	*33 conditions covering:* half of the *Gasch* conditions in the data set all the conditions in the *Spellman* data set the condition of *Cho* data set
	YPR120C **10.01.03** **10.03.01** **10.03.02** YMR199W **10.03.01** YMR179W **11.02.03.04** YML027W **11** YKL113C **10.01.03** **10.01.05.01**	50	*20 genes covering 35 functional categories:* **10** Cell cycle and DNA processing (12: 6.67e-6) **10.01** DNA processing (10: 4.19e-8) **10.01.03** DNA synthesis and replication (8: 9.46e-10) **10.01.05** DNA recombination and DNA repair (5: 3.74e-4) **10.01.05.01** DNA repair (3: 3.34e-3) **10.03** cell cycle (7: 1.71e-3) **10.03.01** mitotic cell cycle and cell cycle control (7: 1.16e-4) **10.03.01.03** cell cycle checkpoints (2: 3.71e-3)	YER095W YGL038C YGR221C YJL074C YMR199W	*9 conditions covering:* all the conditions in the *Spellman* data set excpt "cln2" the condition of *Cho* data set 4 sporadic (none-/ small time-series-experiment) conditions in *Gasch* data

∗ The numbers in the brackets show the number of selected genes in the functional category and the corresponding p-value.

converge. The difference in the scale of ν^{bcl} used for the different seeds is caused by the difference in the noise level (i.e., variation) of the seed genes. From the biological point of view, different choices of ν^{bcl} provide biologists the flexibility in adjusting the trade-off between high sensitivity and high specificity.

Seed 1 mainly consists of genes that are functionally annotated as "cell cycle and DNA processing". Regardless of the input parameters, the three found biclusters are mostly enriched for the same function category.

Seed 2 is experimentally detected (i.e., based on the ChIP chip data) to be regulated by Ndd1, Fkh2 and Mcm1, which are cell cycle regulators. Yet, according to MIPS database, two of the three genes in Seed 2 are annotated as functionally unknown, and the other gene is only associated to "stress response". The results show that the query-driven biclustering algorithm mainly recruited genes that are functionally enriched in categories of "cell cycle and DNA processing" and "cell type differentiation". In addition, the majority of the conditions in any of the found biclusters for Seed 2 are mainly composed of all the cell-cycle-synchronized experiments (from Spellman *et al*. (1998) and Cho *et al*. (1998)). Thus, the biclusters discovered by our algorithm confirm that the three seed genes might have cell cycle related functions.

Seed 3 is composed of 14 genes that are in the functional category of "ribosome biogenesis". The algorithm recruited genes that are highly enriched in the same functional category, especially when a proper ν^{bcl} is used—when $\nu^{\text{bcl}} = 150$, 103 out of the 111 selected genes are found to have the function "ribosome biogenesis"; and when $\nu^{\text{bcl}} = 250$, 85 out of 91 genes are in this functional category. For those genes that are selected for the bicluster but are not associated with "protein synthesis" according to MIPS, we consulted the *Saccharomyces* Genome Database (Balakrishnan *et al*., 2005) and found that all these genes are rather dubious ORFs that overlap with various known ribosomal protein synthesis genes on the other strand of the DNA (see the supplementary information).

Although Seed 3 is obtained by applying the method of De Bie *et al*. (2005) to the data set from Spellman *et al*. (1998), the cell cycle related experimental conditions are seldom selected to be in the bicluster, while almost all the stress response related conditions from Gasch *et al*. (2000) are selected. This result shows that data set from Gasch *et al*. (2000) might be a better data set to look at for the study of "ribosome biogenesis" than those from Spellman *et al*. (1998) and Cho *et al*. (1998), justified by either some biological explanation or the experimental noise presented in the data (see Figure 3C).

Table 1 and 2 also show that the algorithm is able to exclude seed genes that it considers not to belong to the bicluster.

In another experiment (see the supplementary information), we added random genes to the seed genes. The results show that the ability of the algorithm to discover the bicluster in the presence of noisy genes in the seeds depends on the consistency between the (original) seed genes, the deviation of profile of the noisy gene to the mean profile of the (original) seeds, and the number of noisy genes added to the seeds. If the added noisy genes do not contaminate the constructed prior to a large extent, the target bicluster is found, and the noisy genes are excluded. In these cases, the effect of the noisy genes is comparable to that of using a smaller ν^{bcl}.

The likelihood of microarray data is usually highly complex because of the complexity of in the underlying biological process and the non-negligible experimental noise. By introducing a prior, methods based on Bayesian models help to zoom into the local area of interest of the likelihood landscape, and raise the corresponding area in the posterior distribution. Adjustment of the parameter ν^{bcl} has the effect of tuning the zoom. When ν^{bcl} is strong enough, the maximum mode of the posterior distribution provides answer to the query of biologists. The Gibbs sampling procedure produces samples that pictures the posterior distribution as a whole, and consequently the global maximum is decided by Monte Carlo integration of the samples. An alternative way to solve the problem based on the likelihood landscape is by using EM method. However, EM is a method that "climbs" in the posterior distribution and can get stuck in local maxima. To find the global, EM is usually performed for several times and the solution with the greatest posterior probability is chosen. In our experience, the Gibbs sampling procedure turns out to be more efficient to find the global maximum for the query-driven biclustering problem of microarray data, taken into account the massive amount of modes and the fact that it is never easy to decide how many runs of EM can guarantee the discovery of the regional global maximum.

ACKNOWLEDGMENT

This research is supported by research council of K.U.L. (GOA-Mefisto 666, GOA AMBioRICS, IDO, and serval PhD/fellow grants); Flemish government (FWO: PhD/postdoc grants, G.0115.01, G.0240.99, G.0388.03, G.0229.03, G.0241.04, G.0499.04, and research communities; IWO: STWW-Genprom, GBOU-SQUAD, GBOU-ANA); Belgian Federal Science Policy Office (IUAP P5/22); and EU-RTD (ERNSI, FP6-NoE Biopattern; FP6-IP e-Tumours).

REFERENCES

Balakrishnan,R., Christie,K.R., Costanzo,M.C., *et al*. (2005), *Saccharomyces Genome Database*, http://www.yeastgenome.org/.

Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inference of gene changes, *Bioinformatics*, **17**(6), 509-519.

Barash,Y. and Friedman,N. (2002) Context-specific Bayesian clustering for gene expression data, *J. Comput. Biol.*, **9**, 169-191.

Casella,G. and George,E.I. (1992) Explaining the Gibbs sampler, *Am. Stat.*, **46**(3), 167-174.

Cheng,Y. and Church,G.M. (2000) Biclustering of expression data, *ISMB 2000 proceedings*, 93-103.

Cho,R.J., Campbell,M.J., and Winzeler,E.A., *et al*. (1998) A genome-wide transcriptional analysis of mitotic cell cycle, *Mol. Cell* **2**, 65-73.

De Bie,T., Monsieurs,P., Engelen,K., *et al*. (2004) Discovering regulatory modules from heterogeneous information sources, *Proceedings of PSB 2005*.

Gasch,A.P., Spellman,P.T., Kao,C.M., *et al*. (2000) Genomic expression programs in the response of yeast cells to environmental changes, *Mol. Biol. Cell*, **11**, 4241-4257.

Ihmels,J., Friedlander,G., Bergmann,S., *et al*. (2002) Revealing modular organization in the yeast transcriptional network, *Nat. Genet.*, **31**, 370-377.

Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey, *IEEE Transactions on computational biology and bioinformatics*, **1**, 24-45.

McLachlan,G.J., Bean,R.W. and Peel,D. (2002) A mixture model-based approach to the clustering of microarray expression data, *Bioinformatics*, **18**(3), 413–422.

Mewes,H.W., Amid,C., Arnold,R. *et al*. (2004) MIPS: analysis and annotation of protieins from whole genomes, *Nucleic Acids Res.*, **32**, Database issue:D41-D44.

Segal,E., Shapira,M., Regev,A., *et al*., (2003) Module networks: Identifying regulatory modules and their condition-specific regulators for gene expression data, *Nat. Genet.*, **34**(2), 166-176.

Segal,E., Taskar,B., Gasch,A., Friedman,N. and Koller,D. (2001) Rich probabilistic models for gene expression, *Bioinformatics*, **17**(Suppl. 1), S243-S252.

Sheng,Q., Moreau,Y. and De Moor,B. (2003), Biclustering microarray data by Gibbs sampling, *Bioinformatics*, **19**(Suppl. 2), II196-II205.

Spellman,P.T., Sherlock,G., Zhang,M.Q., *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Mol. Biol. Cell*, **9**, 3273-3297.

Tavazoie,S., Hughes,J.D., Campbell,M.J., *et al*. (1999) Systematic determination of genetic network architecture, *Nat. Genet.* **22**, 281-285.

Yeung,K.Y., Fraley,C., Murua,A., *et al*. (2001) Model-based clustering and data transformations for gene expression data, *Bioinformatics*, **17**(10), 977-987.

Table 2. Influence of ν^{bcl} on the biclustering results (continue).

Nr.	Seed genes	ν^{bcl}	Selected genes	Excluded seed genes	Selected conds.
2	YJL051W **99** YGL021W **32.01** YLR190W **99**	10	*54 genes covering 65 functional categories:* **10** Cell cycle and DNA processing (16: 3.91e-3) **10.03** cell cycle (14: 2.52e-4) **10.03.01** mitotic cell cycle and cell cycle control (9: 3.87e-3) **10.03.03** cytokinesis (cell division) / septum formation (4: 2.18e-4) **43** Cell type differentiation (10: 9.68e-4) **43.01** fungal/microorganismic cell type differentiation (10: 9.68e-4) **43.01.03** fungal and other eukaryotic cell type differentiation (10: 9.68e-4) **43.01.03.05** budding, cell polarity and filament formation (2: 2.05e-4)	*none*	*8 conditions:* all the conditions in both *Spellman* and *Cho* data set except for "cln3" 3 sporadic conditions in *Gasch* data set
		20	*24 genes covering 34 functional categories:* **10** Cell cycle and DNA processing (10: 1.62e-3) **10.03** cell cycle (10: 2.93e-5) **10.03.01** mitotic cell cycle and cell cycle control (6: 2.47e-4) **10.03.03** cytokinesis (cell division) / septum formation (4: 9.11e-6) **43** Cell type differentiation (6: 2.367e-3) **43.01** fungal/microorganismic cell type differentiation (6: 2.367e-3) **43.01.03** fungal and other eukaryotic cell type differentiation (6: 2.367e-3) **43.01.03.05** budding, cell polarity and filament formation (6: 2.82e-4)	*none*	*11 conditions:* all the conditions in both *Spellman* and *Cho* data set 4 sporadic conditions and "nitrogen depletion" in *Gasch* data set
		30	*16 genes covering 32 functional categories:* **10** Cell cycle and DNA processing (6: 2.74e-2) **10.03** cell cycle (6: 2.65e-3) **10.03.03** cytokinesis (cell division) / septum formation (3: 9.99e-5) **43** Cell type differentiation (4: 1.39e-2) **43.01** fungal/microorganismic cell type differentiation(4: 1.39e-2) **43.01.03** fungal and other eukaryotic cell type differentiation(4: 1.39e-2) **43.01.03.05** budding, cell polarity and filament formation (4: 3.35e-3)	*none*	*5 conditions* all the conditions in both *Spellman* and *Cho* data set except for "cln3"
3	YGR148C **12.01** YGL189C **12.01** YER056C-A **12.01** YER131W **12.01** YGL031C **12.01** YGL103W **12.01** YER102W **12.01** YLR167W **12.01** **14.13.01**	100	*878 genes in 147 functional categories:* **01.03** nucleotide metabolism (44: 5.15e-7) **01.03.01** purine nucleotide metabolism (18: 2.00e-5) **01.03.04** pyrimidine nucleotide metabolism (14: 2.11e-5) **12** Protein synthesis (204: 8.76e-12) **12.01** ribosome biogenesis (138: 1.28e-11) **12.04** translation (42: 8.95e-12) **12.10** aminoacyl-tRNA-synthetases (19: 1.03e-17) **11.02.01** rRNA synthesis (23: 2.05e-4) **11.02.02** tRNA synthesis (15: 6.10e-6) **11.04** RNA processing (73: 3.36e-6) **11.04.01** rRNA processing (55: 9.22e-12) **16.03.03** RNA binding (5: 3.14e-4)	*none*	*61 conditions:* 60 conditions from the *Gasch* data set "size-based synchronization" (*Spellman*)
	YLR029C **12.01** YLR333C **12.01** YOL127W **12.01**	150	*111 genes covering 18 functional categories:* **12** Protein synthesis (104: 7.85e-12) **12.01** ribosome biogenesis (103: 1.02e-11)	YLR333C	*60 conditions:* 59 conditions from the *Gasch* data set "size-based synchronization" (*Spellman*)
	YOL040C **12.01** YLR344W **12.01** YLR441C **12.01** **34.11.03.07**	250	*91 genes covering 15 functional categories:* **12** Protein synthesis (85: 3.33e-12) **12.01** ribosome biogenesis (85: 1.20e-11)	YLR333C YOL040C YOL127W	*45 conditions:* all from the *Gasch* data set