Gene expression

Query-driven module discovery in microarray data

Thomas Dhollander[1*], Qizheng Sheng[1], Karen Lemmens[1], Bart De Moor[1], Kathleen Marchal[1,2] and Yves Moreau[1]

[1]Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, Leuven-Heverlee, 3001, Belgium, [2]CMPG, Department of Microbial and Molecular Systems, K.U.Leuven, Kasteelpark Arenberg 20, B-3001 Leuven, Belgium

ABSTRACT

Motivation: Existing (bi)clustering methods for microarray data analysis often do not answer the specific questions of interest to a biologist. Such specific questions could be derived from other information sources, including expert prior knowledge. More specifically, given a set of seed genes which are believed to have a common function, we would like to recruit genes with similar expression profiles as the seed genes in a significant subset of experimental conditions.

Results: We propose a novel Bayesian query-driven biclustering framework in which the prior distributions allow introducing knowledge from a set of seed genes (query) to guide the pattern search. In two well-known yeast compendia, we grow highly functionally enriched biclusters from small sets of seed genes using a resolution sweep approach. In addition, relevant conditions are identified and modularity of the biclusters is demonstrated, including the discovery of overlapping modules. Finally, our method is also robust against incorrect seed genes, deals with missing values naturally and performs well on a recent artificial data benchmark study.

Availability: Software is available on the supplementary website.

Contact: Thomas.dhollander@esat.kuleuven.be

Supplementary Information: Available on http://homes.esat.kuleuven.be/~tdhollan/Supplementary_Information_Dhollander_2007/index.html

1 INTRODUCTION

Suppose biologists have at hand a specific gene or set of genes which they know or expect to be related to some common biological function. We refer to these genes as the seed genes hereafter. In this context, it is often of interest to be able to use available high throughput data to recruit additional genes that are involved in this function. More specifically, we have in mind various questions or *queries* such as "which genes involved in a specific protein complex are coexpressed?" or "given a set of known disease genes, how to select new candidate genes that might be linked to the same disease?".

Although we aim at a more general discussion on query-driven module discovery, our analyses and models are tuned towards the use of microarray data. Our contribution is mainly based on three observations:

- Bayesian probabilistic models have shown promise in providing answers to such specific questions or queries, by transforming the existing knowledge of biologists into prior probability distributions that are incorporated into the model (see, for example, Gevaert *et al.* (2006) and Bernard and Hartemink (2005)). Probability distributions usually contain many modes, due to the complexity of the underlying biological process. Because in Bayesian models the likelihood of the data is multiplied by the prior to obtain the posterior probabilistic model, a proper prior can substantially raise the mode that is most relevant to the biological question in the posterior distribution: the prior is used to zoom in on a specific part of the likelihood. By contrast, many well-known clustering methods suffer a lack of sharpness that has prevented them from surpassing their rather vague exploratory role.
- Probabilistic models have become a popular choice for modeling high throughput genomic data (see eg. Friedman (2004)), because they allow natural handling of high noise levels.
- Current microarray compendia consist of measurements in multiple biological conditions and it may not be clear which conditions are truly most relevant to the biological question at hand. Therefore, simultaneous identification of the appropriate subset of experimental conditions (features), often referred to as biclustering (Madeira and Oliveira, 2004), has become a profitable extension to classical cluster analysis. In other words, we want to take into account the fact that some genes are only tightly coexpressed in a subset of experimental conditions (for these experimental conditions, their regulatory program significantly overlaps). Classical clustering cannot always recover such sets of genes if they are obscured by a large set of irrelevant conditions (see, for example, Prelic *et al.* (2006)). Moreover, the discovery of relationships between the genes and the conditions may provide important information for unveiling genetic pathways (Van den Bulcke *et al.*, 2006).

These observations led to the development of a Bayesian probabilistic framework. The last observation incited us to generalize the query further (from clustering to biclustering) to include a selection of appropriate features: "which genes are functionally related to the seed genes and which features are relevant for this biological function?" We will use the terms 'bicluster' or 'module' to indicate such a set of genes and its relevant conditions.

Biclustering techniques for microarray data (see Madeira and Oliveira (2004) for a survey) are recently receiving increasing attention in bioinformatics (e.g., Gibbs biclustering (Sheng *et al.*, 2003), OPSM (Ben-Dor *et al.*, 2003), Samba (Tanay *et al.*, 2002), ISA (Bergmann *et al.*, 2003), xMotif (Murali and Kasif, 2003), CC (Cheng and Church, 2000)). However, most existing methods reveal global patterns in the data, rather than patterns around a specific biologically inspired query. Only few existing algorithms, such as the (Iterative) Signature Algorithm (Bergmann *et al.*, 2003) and the Gene Expression Mining Server (Wu and Kasif, 2005) allow more directed searches, but they lack a clear-cut statistical interpretation. Moreover, they heavily depend on the choice of (at least) two parameters to specify the desired resolution.

*To whom correspondence should be addressed.

© Oxford University Press 2007

In the remainder of the paper we describe the model and discuss how Gibbs sampling (Gelman *et al.*, 2004) or Expectation Maximization (Dempster *et al.*, 1977) can be used for model estimation. Furthermore, we study the performance on a series of artificial data sets taken from Prelic *et al.* (2006). Finally, we show results obtained on the combined Gasch *et al.* (2000) and Spellman *et al.* (1998) microarray gene expression data sets.

2 SYSTEMS AND METHODS

2.1 Artificial data

The artificial data were taken from the supplementary website of Prelic *et al.* (2006). In a first scenario (S1), the data consist of 10 binary modules (expression value 1) embedded in a zero background (expression value 0). This simple setup is complicated by adding Gaussian noise (A) or allowing module overlap (B). In (A), the modules have 10 genes and 5 conditions, while the modules in (B) each contain 10 + *k* genes and 10 + *k* conditions, where *k* is the number of genes and conditions in common between overlapping modules. A second scenario (S2) describes a similar case (same module sizes), only this time the data is not binary but continuous. The background values are samples from a Gaussian distribution, the bicluster values in each column are equal (B) or equal up to some Gaussian noise (A). For details, we refer to Prelic *et al.* (2006). No preprocessing or postprocessing (such as filtering) was performed.

In all experiments, the seed consisted of a single gene correctly belonging to one of 10 artificial modules. We repeated the biclustering process 10 times, each time with a randomly selected seed gene from a different module. The resulting biclusters were then scored with *module recovery* and *bicluster relevance* scores as described in Prelic *et al.* (2006) and in Supplementary File 1. The *module recovery* score indicates how well the gene content of the 'ideal' modules is on average reflected in the (best matching bicluster in the) bicluster results. The *bicluster relevance* score is related to the relevance of the set of modules in the output. Both scores are maximal and equal to one if both module sets are equal.

2.2 Resolution sweep

For the resolution sweep in artificial data, we linearly increased the variance prior from 0 to 3 times the background variance over 300 iterations. On the real data, we used a sweep from 0 to the background variance, over 500 iterations. We did not increase the variance further, since most modules already contained on the order of 1000 genes and represented very general functional categories when the bicluster variance equaled the background variance. The variance priors were chosen to be very informative (v^{bcl} = 1e10 ≈ ∞). To automatically detect the resolutions of interest, we identified the local maxima in the likelihood $p(\mathbf{X} \mid \mathbf{g}, \mathbf{c}, \boldsymbol{\theta})$ on the resolution sweep path. Functional enrichment scores were calculated for the corresponding modules.

2.3 Initialization

For seeds containing more than one gene, we add a dummy gene to the data set, with a profile equal to the mean profile of the seed genes. This yields an improved initialization for the resolution sweep (start with one dummy gene and all conditions, rather than no genes and no conditions). Similarly, when running the algorithm with a fixed variance prior, we initialize all condition labels to 1. In all cases, the label of the dummy gene (or the seed gene, if only one seed gene is used) is initially set to 1 while all other gene labels equal 0.

2.4 Robustness

The noisiest data set (noise setting 0.25) of scenario S1A was used to assess robustness of the biclustering result against incorrect seed genes. The total number of seed genes was fixed to 10, but a growing fraction of it was sampled from a pool of random profiles, with expression values generated using the overall mean and standard deviation of the data. Again, the default resolution sweep approach with maximally informative priors (1e10 prior observations) was adopted.

2.5 Seed genes for combined Gasch and Spellman data set

Seeds were taken from the supplementary website of one of our recent publications on module discovery in yeast (Lemmens *et al.*, 2006). They correspond to very small gene sets that are selected based on similarity in motif, expression and ChIP-chip data using the Apriori algorithm (Agrawal and Imielenski, 1993). For more details, see Lemmens *et al.* (2006).

2.6 Yeast data

We downloaded the yeast data from the supplementary material of Gasch *et al.* (2000) and Spellman *et al.* (1998). As in Lemmens *et al.* (2006), we normalized the log ratios of both data sets per gene (subtracting the mean of each profile and dividing by the standard deviation). No further preprocessing was carried out.

2.7 Functional enrichment

The hypergeometric distribution was used to determine which Gene Ontology Biological Process categories (Ashburner *et al.*, 2000) were statistically overrepresented in the selected biclusters resulting from the resolution sweep approach (Sokal and Rohlf, 1995). A cutoff of 0.5 on the label probabilities was used to determine gene and condition cluster membership. All known GO-BP labels from Ensembl (Hubbard *et al.*, 2007) were propagated towards the root of the hierarchy. There was no correction for multiple testing.

3 MODEL AND ALGORITHMS

The core of the probabilistic framework resembles that of Sheng *et al.* (2003), the main ingredients being a column-wise statistical model for the bicluster expression (indicated by the superscript 'bcl'), a column-wise statistical model for the background expression (superscript 'bgd'), and hidden labels for the genes (**g**) and conditions (**c**) to indicate bicluster membership. Each column *j* of the (*n* x *m*) expression data matrix **X** represents an experimental condition, each row *i* a gene. In Sheng *et al.* (2003), we applied a Bayesian hierarchical model for discretized data with a multinomial likelihood and a Dirichlet prior to the problem of biclustering patients (finding clusters of patients that have similar expression patterns over a subset of their genes). In this paper we improve and adjust the framework to tackle the dual problem, in which we bicluster genes (looking for genes with similar profiles, rather than patients with similar gene expression). Supplementary Figure 1 shows a conceptual scheme of the framework.

For column-wise (condition-wise) Gaussian distributions with parameters $\theta_j = (\mu_j, \sigma_j)$, the conjugate Normal - Inverse χ^2 priors are given by

$$p(\mu_j, \sigma_j) = p(\mu_j \mid \sigma_j)\, p(\sigma_j) \propto \mathrm{N} - \mathrm{Inv} - \chi^2.$$

In Gelman *et al.* (2004), this type of prior is discussed and motivated in detail. If x_{ij} indicates the expression level of gene *i* in experimental condition *j*, the corresponding distributions for bicluster and background can be written as

$$\begin{cases} p^{\,\mathrm{bcl}}(x_{ij}) \propto N\left(\mu_j^{\mathrm{bcl}}, (\sigma_j^{\mathrm{bcl}})^2\right) \\ p(\mu_j^{\mathrm{bcl}} \mid \sigma_j^{bcl}) \propto N\left(\varphi_j^{\mathrm{bcl}}, \dfrac{(\sigma_j^{\mathrm{bcl}})^2}{\kappa^{\mathrm{bcl}}}\right) \\ p\left((\sigma_j^{\mathrm{bcl}})^2\right) \propto \mathrm{Inv} - \chi^2\left(\nu^{\mathrm{bcl}}, (s_j^{\mathrm{bcl}})^2\right) \end{cases} \quad \begin{cases} p^{\,\mathrm{bgd}}(x_{ij}) \propto N\left(\mu_j^{\mathrm{bgd}}, (\sigma_j^{\mathrm{bgd}})^2\right) \\ p(\mu_j^{\mathrm{bgd}} \mid \sigma_j^{\mathrm{bgd}}) \propto N\left(\varphi_j^{\mathrm{bgd}}, \dfrac{(\sigma_j^{\mathrm{bgd}})^2}{\kappa^{\mathrm{bgd}}}\right) \\ p\left((\sigma_j^{\mathrm{bgd}})^2\right) \propto \mathrm{Inv} - \chi^2\left(\nu^{\mathrm{bgd}}, (s_j^{\mathrm{bgd}})^2\right) \end{cases}$$

Of course, we do not know in advance which expression values x_{ij} should be captured by the bicluster model; the complete model depends on the hidden gene (**g**) and condition labels (**c**). Indeed, we are mostly interested in the marginal probability distributions (or at least in the corresponding

marginal means) of these hidden gene labels $p(g_i \mid \mathbf{X})$ and condition labels $p(c_j \mid \mathbf{X})$, but it is intractable to describe these distributions analytically. To overcome this problem, strategies like Gibbs sampling (Gelman *et al.*, 2004) or Expectation-Maximization (EM) (Dempster *et al.*, 1977) can be applied to explore the full posterior and marginal distributions using samples from the full conditional probability distributions only. For the gene labels **g**, these conditionals are Bernouilli distributions:

$$p(g_i = 1 \mid \mathbf{g}_{\neq i}, \mathbf{c}, \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{X}) \propto \text{Bern}(\alpha_i)$$

$$\text{with} \quad \frac{\alpha_i}{1-\alpha_i} = \left(\prod_{j \in \text{bcl}} \frac{p^{\text{bcl}}(x_{ij})}{p^{\text{bgd},1}(x_{ij})} \right) \left(\frac{\xi_{g1} + \|\mathbf{g}_{\neq i}\|_1}{\xi_{g0} + n - 1 - \|\mathbf{g}_{\neq i}\|_1} \right).$$

In this expression, $B(\xi_{g1}, \xi_{g0})$ represents the Beta prior distribution on the probability that a gene belongs to the bicluster, n the total number of genes, $\boldsymbol{\theta}$ the set of model parameters (μ, σ), $\boldsymbol{\psi}$ the total set of hyperparameters (κ, ν, s, φ, ξ_0, ξ_1) and $\|\mathbf{g}_{\neq i}\|_1$ the one norm of the current (binary) gene label vector, except for label i. In other words, the second factor depends on the number of genes currently in the bicluster as well as the prior knowledge on the bicluster size. It is equal to one when the prior probability that a gene belongs to the bicluster is assumed to be known and equal to 50% ($\xi_{g1} = \xi_{g0} = \infty$). The first factor corresponds to the likelihood ratio of bicluster versus background model.

A similar model holds for the full conditional distribution of the condition labels **c**:

$$p(c_j = 1 \mid \mathbf{g}, \mathbf{c}_{\neq j}, \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{X}) \propto \text{Bern}(\beta_j)$$

$$\text{with} \quad \frac{\beta_j}{1-\beta_j} = \left(\prod_{i \in \text{bcl}} \frac{p^{\text{bcl}}(x_{ij})}{p^{\text{bgd},0}(x_{ij})} \right) \left(\prod_{i \in \text{bgd}} \frac{p^{\text{bgd},1}(x_{ij})}{p^{\text{bgd},0}(x_{ij})} \right) \left(\frac{\xi_{c1} + \|\mathbf{c}_{\neq j}\|_1}{\xi_{c0} + m - 1 - \|\mathbf{c}_{\neq j}\|_1} \right).$$

Again, a Beta prior allows introducing prior knowledge on the bicluster size (last factor between brackets). The first factor between brackets is similar to the first factor in the expression for the gene label conditionals. However, an additional second likelihood term appears. Indeed, the likelihood ratio for the genes in column j is in general different from one, even if a gene is part of the background. If condition label j is zero, the entire j-th column is captured by one background model (bgd,0). On the other hand, if the j-th column label equals one, the background model that describes the expression of those background genes should only be constructed with the background genes (bgd,1). So, although we keep the model parameters fixed here, strictly speaking there are two background models rather than one single model. In most cases, however, this distinction is primarily a technical issue and the difference between bgd,0 and bgd,1 only matters when very large biclusters occur. For more details, we refer to Supplementary File 1 or the website.

Given the choice of the Normal – Inv – χ^2 priors, the full conditional distributions for the model parameters are given by

$$p\left((\sigma_j^{\text{bcl}})^2 \mid \mathbf{g}, \mathbf{c}, \boldsymbol{\theta}_{\neq \sigma_j^{\text{bcl}}}, \boldsymbol{\psi}\right) = p\left((\sigma_j^{\text{bcl}})^2 \mid \mathbf{g}, c_j, \nu^{\text{bcl}}, (s_j^{\text{bcl}})^2\right) \propto \text{Inv} - \chi^2\left(\eta_j^{\text{bcl}}, (\varsigma_j^{\text{bcl}})^2\right)$$

$$\begin{cases} \eta_j^{\text{bcl}} = \nu^{\text{bcl}} + \|\mathbf{g}\|_1 \\ (\varsigma_j^{\text{bcl}})^2 = \dfrac{\nu^{\text{bcl}} (s_j^{\text{bcl}})^2 + \|\mathbf{g}\|_1 (\bar{\sigma}_j^{\text{bcl}})^2}{\nu^{\text{bcl}} + \|\mathbf{g}\|_1} \end{cases}$$

and

$$p(\mu_j^{\text{bcl}} \mid \mathbf{g}, \mathbf{c}, \boldsymbol{\theta}_{\neq \mu_j^{\text{bcl}}}, \boldsymbol{\psi}) = p(\mu_j^{\text{bcl}} \mid \mathbf{g}, c_j, \sigma_j^{\text{bcl}}, \varphi_j^{\text{bcl}}, \kappa^{\text{bcl}}) \propto N\left(\gamma_j^{\text{bcl}}, (\lambda_j^{\text{bcl}})^2\right)$$

$$\begin{cases} \gamma_j^{\text{bcl}} = \dfrac{\dfrac{\varphi_j^{\text{bcl}}}{(\tau_j^{\text{bcl}})^2} + \dfrac{\bar{\mu}_j^{\text{bcl}}}{\frac{(\sigma_j^{\text{bcl}})^2}{\|\mathbf{g}\|_1}}}{\dfrac{1}{(\tau_j^{\text{bcl}})^2} + \dfrac{1}{\frac{(\sigma_j^{\text{bcl}})^2}{\|\mathbf{g}\|_1}}} = \dfrac{\kappa^{\text{bcl}} \varphi_j^{\text{bcl}} + \|\mathbf{g}\|_1 \bar{\mu}_j^{\text{bcl}}}{\kappa^{\text{bcl}} + \|\mathbf{g}\|_1} \\ (\lambda_j^{\text{bcl}})^2 = \dfrac{1}{\dfrac{1}{(\tau_j^{\text{bcl}})^2} + \dfrac{1}{\frac{(\sigma_j^{\text{bcl}})^2}{\|\mathbf{g}\|_1}}} = \dfrac{(\sigma_j^{\text{bcl}})^2}{\kappa^{\text{bcl}} + \|\mathbf{g}\|_1} \end{cases}$$

The prior parameters κ and ν can be interpreted as 'pseudocounts' or the number of 'prior observations' for the estimates of the mean and variance respectively. The resulting estimates for means (variances) are weighted means of the observed sample mean (variance) and the prior mean (variance). For brevity, we omitted the formulas for the background model parameters, which are similar.

Gibbs sampling consists of sampling alternatively from the full conditional distributions mentioned above. After convergence, we easily obtain (correlated) samples from the wanted marginal gene and condition labels probabilities.

As an alternative to the Gibbs sampling approach, we consider using EM to fit the model parameters (Dempster *et al.*, 1977). Instead of sampling 0-1 labels from the Bernouilli label probability distributions, the expected label probabilities are used as weights. The EM variant corresponds to a gradient ascent to a local mode in the posterior density (Gelman *et al.*, 2004).

To illustrate the advantages of biclustering over classical clustering, we included a clustering variant of the EM algorithm as well. The model used here is essentially identical but no condition sampling or weighting was applied.

4 IMPLEMENTATION

All algorithms were implemented in the R language and environment for statistical computing (R Development Core Team, 2006). To download the GO Biological Process functional annotations, we used packages from the Bioconductor repository (Gentleman *et al.*, 2004).

5 RESULTS AND DISCUSSION

5.1 Parameter interpretation

Before tying up any parameters, it is useful to stress the flexibility of the presented framework. The remainder of this paper describes a particular instantiation of the model, using specific choices for the distributions and their priors, which we deemed most intuitive. One should keep in mind that alternative ways to introduce the prior knowledge exist.

5.1.1 User-defined parameters

Strictly speaking, 4 groups of parameters specify the priors:

(1) The parameters for the prior on the mean and variance of the bicluster model

(2) The parameters for the prior on the mean and variance of the background model

(3) Two parameters for the Beta prior $B(\xi_{g1}, \xi_{g0})$ on the prior probability that a gene belongs to the bicluster

(4) Two parameters for the Beta prior $B(\xi_{c1}, \xi_{c0})$ on the prior probability that an experimental condition belongs to the bicluster

Although these parameters all have a clear-cut statistical interpretation, it is not desirable to choose all of them manually. Therefore, we assume that the prior probability of a gene (condition) belonging to the bicluster is unknown. The corresponding Beta priors can then be fixed to B(1,1) (noninformative, uniform distribution), eliminating (3) and (4). In other words, we assume that no prior knowledge about the bicluster size is available. To sufficiently restrict the freedom of the bicluster (preventing it to drift too far away from the seed profile) and to limit the number of user parameters, we force the mean of the bicluster to be equal to the mean of the seed genes ($\varphi_j^{bcl} = \bar{\mu}_j^{seed}$ and $\kappa^{bcl} = \infty$) in the selected conditions by default (tying up additional parameters in (1)).

We do not make the background prior parameters (2) available to the user. Indeed, any background of interest contains many genes and hence the likelihood (of the background data, including all genes except for those that are currently in the bicluster) and the prior (constructed with all genes except for the seed genes) usually agree. Therefore, in most cases the strength of these priors is rather unimportant. However, to avoid those cases in which very large modules push away the background distribution, we require the background statistics to remain close to the expected ones (equal in terms of means, similar in terms of variance) by setting the number of prior observations for the background variance (ν^{bgd}) equal to the total number of genes by default. The outcome of the algorithm is robust against the precise choice of these background parameters (and more robust if the module signal is stronger).

In other words, we chose to work with the prior on the bicluster variance primarily. As explained in the Model and Algorithms section, this prior is a (scaled) inverse chi-square distribution with two kinds of parameters: s_j^{bcl} indicates the prior variance while ν^{bcl} refers to the number of prior observations (seed strength). For a bicluster containing 10 genes at a certain point in the sampling procedure, 90 prior observations would mean that the resulting variance of the bicluster is sampled from a distribution that is determined for 90% by the prior variance and for 10% by the variance of the 10 genes in the bicluster.

5.1.2 Noninformative variance priors

In most artificial data scenarios, results were only weakly dependent on the choice of the prior parameters ν^{bcl} and s_j^{bcl}. In general, priors with low information content (weak priors) performed well here, since the patterns were very strong in most cases and this strength was (almost) equal for all modules. Therefore, it is in principle possible to use a noninformative prior for the variance (small number of prior observations ν^{bcl} with large variance s_j^{bcl}, for instance equal to background variance) and still detect most of the modules (data not shown). In this case, we do not assume any prior knowledge on the size of the module, neither in terms of the prior probability of a gene (condition) belonging to the bicluster, nor in terms of prior variance. With these settings, the algorithm becomes almost parameterless up to the choice of the probability distribution type.

5.1.3 Resolution sweep

In real data sets, we expect the choice of the number of prior observations to be more crucial. Indeed, these data are typically dominated by a small number of very strong biclusters (such as ribosome biogenesis in yeast). Therefore, stronger priors are needed to extract interesting but smaller (statistically less significant) patterns around the seed. For example, simulations with noninformative or seed-based variance priors on the Gasch *et al.* (2000) and Spellman *et al.* (1998) yeast expression data showed that the algorithm converged to a part of the dominant ribosome biogenesis module in many cases. Indeed, the statistical significance of very strong correlation between small numbers of genes in a large number of experimental conditions might be exceeded by weaker correlation between larger numbers of genes over a smaller subset of the conditions. Hence, it may happen that the profiles of the seed genes match better with the ribosome biogenesis expression pattern than with the background pattern over a sufficiently large number of conditions, causing the algorithm to get stuck in the ribosome biogenesis mode of the posterior distribution (see also Section 5.4.1). Therefore, it seems preferable to explicitly zoom in on the appropriate peak by using an informative variance prior. In general, a decrease in prior variance will give rise to smaller modules.

Since the most interesting setting for the prior variance is unknown (and noninformative variance priors do not work well in practice), it seems necessary to explicitly test a whole range of prior settings. For every one of those settings, we have to wait until local convergence occurs. Unfortunately, a large number of settings may result in very large or very small (less interesting) modules. To quickly identify the parameter ranges of interest, we therefore suggest a *resolution sweep* approach (see Systems and Methods, resolution sweep) in which the prior variance is slowly (and linearly) increased while the Gibbs sampler or EM algorithm is running. In fact, this means that the starting point at each prior setting is equal to the mode that was found with a slightly smaller value of the variance prior. Because we start close to the seed (Systems and Methods section, initialization) and stay in a mode of interest at any time, the algorithms remain near convergence, and the sweep requires a much smaller total number of iterations. In our experience, the results obtained with both types of multi-resolution analyses were similar in this query-based setting, if not the same (data not shown). Of course, the sceptical reader may use the latter sweep as an indication only, and repeat the experiment with a parameter value chosen from a region of interest.

Varying prior variance means traveling through the modular structure of the data in the neighborhood of the seed. Linear increases of the prior parameter result in discrete steps for the observed module sizes. In Section 5.4.1, we report an example that evolves from very specific cell-cycle related functions over less specific ones to ribosome biogenesis related function (when the prior variance is large). The corresponding figures (and many more examples on the supplementary website) clearly reveal some well-known modularity properties of genetic regulatory networks (Ihmels *et al.*, 2002).

All artificial and real data simulations shown here were obtained using this resolution sweep approach. From the results, we extracted the most promising modules by searching for local optima in the likelihood.

5.2 Gibbs sampling versus EM

Gibbs samplers are guaranteed to converge eventually (see, eg. Gelman *et al.* (2004)). This means that independent of the initialization, after a so-called *burn-in period* the obtained samples are (dependent) samples from the full posterior distribution. However, it is well-known that very many samples may be needed for high dimensional posteriors with highly correlated parameters (Gelman *et al.*, 2004). Therefore, we cannot always expect global convergence within a reasonable amount of time. In some sense, using a Gibbs sampler in such scenarios is like performing a stochastic gradient search in the posterior landscape. The sampler may still allow overcoming small ridges in the probability surface, but is practically unable to move to distinct regions of probability space.

To overcome this problem, repeated initialization of the Gibbs chain can be used (possibly with masking of the obtained biclusters, as in Sheng *et al.* (2003)). Similar restarting procedures are often applied in a K-means or EM clustering context, where convergence is always local.

In the query-driven context under study, local convergence is often sufficient because

- the query is introduced through strong priors on the model parameters (although the priors on the bicluster size are noninformative, see below). Hence, priors are used to zoom in on the likelihood (see Systems and Methods) and many modes will no longer be of interest.
- the knowledge represented by the seed genes can be used for clever initialization (see Systems and Methods).

Although appropriate sampling procedures could allow a more detailed exploration of possibly multi-modal posteriors, we will not further discuss the Gibbs sampling alternative but focus on EM in the remainder of this paper. The latter approach does not only allow faster implementations, it also eases the interpretation of the resolution sweep results: when small variations of the prior variance result in abrupt changes of the module content, these changes are guaranteed to represent qualitative changes in the posterior landscape. By contrast, in a Gibbs sampling approach, the mobility of the sampling chain itself could yield similar effects, complicating the interpretation. For similar reasons, EM allows using local optima in the likelihood to select modules of interest (see Systems and Methods section, resolution sweep).

5.3 Artificial expression data

Following the approach in Prelic *et al.* (2006), we systematically evaluated our algorithm on several artificial data sets in two scenarios S1 and S2, containing noiseless overlapping modules (A) and noisy non-overlapping modules (B). More details on the setup and the resolution sweep settings can be found in the Systems and Methods section. Figure 1 is comparable with the module recovery plots in Prelic *et al.* (2006), since our input data were collected from their supplementary website and the performance measures are equal to those used in that paper. Due to a lack of space, we moved the corresponding bicluster relevance plots to our supplementary material (Supplementary Figure 2 or the website).

Figure 1 clearly illustrates that query-based EM biclustering performs very well on these data, even though the setup uses modules of equal strength and therefore does not fully exploit the advantage of the query-based setting. For every scenario, scores are comparable to the best scores in Prelic *et al.* (2006). The performance of the clustering variant is acceptable (better than the clustering results described in Prelic *et al.* (2006)) but clearly less good than that of the biclustering variant, especially in noisy situations (S1A and S2A). Moreover, it does not allow identifying experimental conditions associated with the biological function under study.

5.3.1 Robustness against incorrect seed genes

Because our method relies on sets of seed genes, it is important to examine its resistance against impurities in the seed input. Figure 2 shows the performance of the EM approach with resolution sweep as a function of the percentage of incorrect seed genes for the noisiest setting in scenario S1A. Even for the relatively high noisy level in this data set, performance of the biclustering algorithm remains good for impurities up to 70%. The EM clustering variant is less robust, as illustrated by a decrease in performance for impurities over 40% (Figure 2). Indeed, biclustering excludes those conditions in which the noisy genes disturb the pattern too strongly, leading to superior robustness properties. Nevertheless, we feel that the quality of the seed is in general more important than its size. On the artificial data set, less informative priors (fewer prior observations) were usually more robust; the trade-off involved here was discussed in some detail in Sections 5.1.2 and 5.1.3.

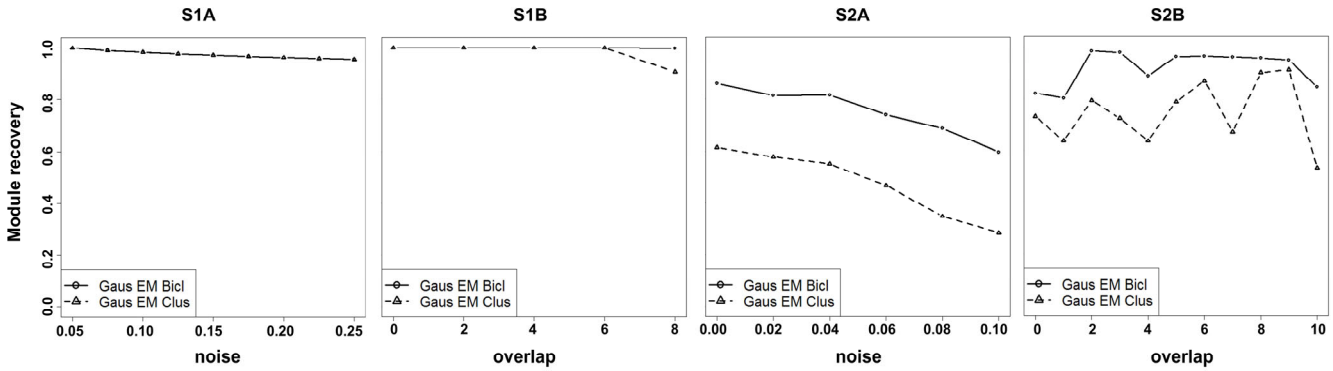


Figure 1: Evolution of module recovery scores as a function of noise (indicated by the letter A) and overlap (letter B), in two artificial data scenarios (S1 and S2), taken from (Prelic *et al.*, 2006). The EM bicluster variant is shown as a solid line, the EM cluster variant as a dashed line.

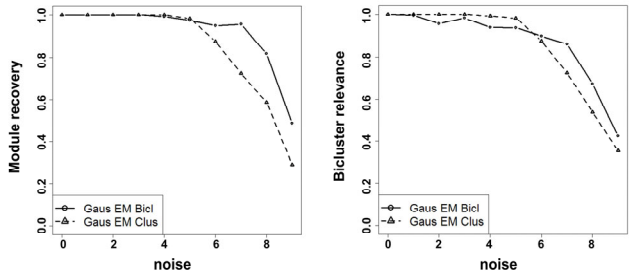


Figure 2: Robustness properties of the EM variants for clustering and biclustering on scenario S1A, with default resolution sweep settings. The x-axis represents the number of incorrect seed genes, on a total of 10 seed genes.

5.3.2 Shortcomings of artificial data analysis

Although artificial data are helpful to gain understanding in properties of algorithms, creating realistic scenarios and choosing good performance measures remains extremely tricky (Supplementary File 2). As mentioned before, we decided to use artificial data from a recent comparative study (Prelic et al., 2006) in order to avoid any bias in creating our own data sets.

5.4 Yeast expression data

We evaluated the performance of our approach by applying the EM biclustering version to a concatenation of two well-known yeast expression compendia (Gasch *et al.*, 2000; Spellman *et al.*, 1998). Seeds were taken from the supplementary website of a recent paper on module discovery in yeast (Lemmens *et al.*, 2006). The query-driven biclustering algorithm can act as a more sophisticated approach to the so-called seed extension step in (Lemmens *et al.*, 2006). Note that we use a combined data set, instead of the separate treatment in Lemmens *et al.* (2006). For instance, we also consider the Spellman *et al.* (1998) data when extending the seeds identified in the Gasch *et al.* (2000) data.

The results of our analyses of over 100 sets of seed genes can be found on our supplementary website. In most cases, we could find highly enriched biclusters associated with functions similar to those described in Lemmens *et al.* (2006). Additionally, we gain information through condition selection and reveal relationships between functions. Moreover, the suggested approach is robust against noise. Most *cell cycle* seeds ultimately evolve into ribosome biogenesis related modules, while most *nutrient-deprived* seeds evolve over *nitrogen compound metabolism* into *aerobic respiration* and more general *energy-related* functions. For g*alactose metabolism* seeds we did not observe any function changes over the tested resolution range.

5.4.1 A cell cycle bicluster example

We found several examples showing overlapping biclusters, one of which is shown in Figure 3 and Figure 4. The seed consists of two genes and was obtained using the Spellman data set (together with ChIP-chip and motif data, as discussed in Lemmens *et al.* (2006)). Figure 3 illustrates how the bicluster grows (in size) when the (prior) variance is gradually increased (resolution sweep). In the first selected module (see Systems and Methods), only the dummy gene is included, together with all conditions. Upon increasing the prior variance, the number of selected conditions slowly decreases while the number of genes starts to increase. The most overrepresented functional category for the second selected module is DNA-dependent DNA replication (p = 4.47e-2). When we further increase the variance, the algorithm picks up the signal from stronger mitosis (p = 2.33e-4) and cell division (p = 3.73e-8) modules. Around iteration 330, there is an abrupt transition to the dominant ribosomal module (p < 1e-16). The latter change (from module **A** to **B** in Figure 3) includes a significant drop in the number of selected conditions, together with an increase in the number of selected genes. Notably, many Spellman *et al.* (1998) cell cycle conditions are lost, in agreement with the change in function.

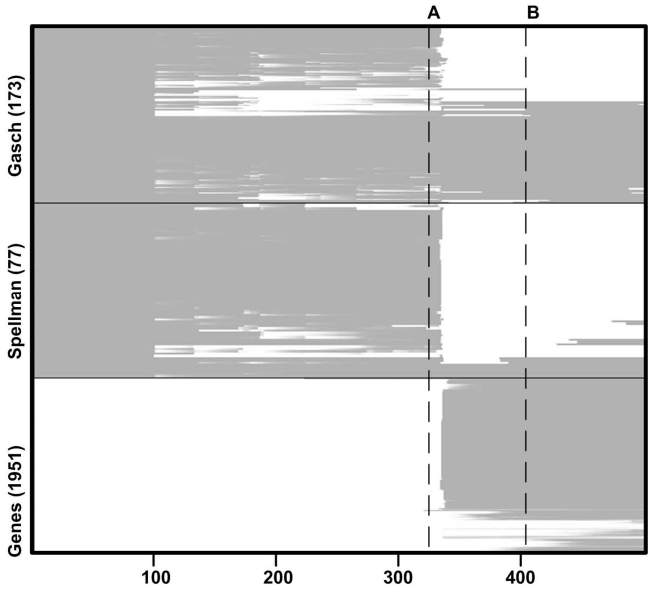


Figure 3: Evolution of bicluster S9 (based on (Spellman *et al.*, 1998) seed number 9) with increasing variance from left to right. Ten modules at various resolutions are automatically detected as interesting (see Systems and Methods). Functional enriched GO Biological processes for these four modules change from DNA-dependent DNA replication (p = 4.47e-2) over mitosis (p = 2.33e-4) to cell division (**A:** p = 3.73e-8) and ribosome biogenesis related functions (**B:** p < 1e-16), as the prior variance increases. Gray indicates high gene or condition scores. Conditions and genes were clustered based on their score vectors, to improve the visualization. For clarity, only genes for which the sum of the scores over all iterations exceeded 1 have been included.

The described transition is interesting because it reveals overlapping bicluster patterns in the data. This is illustrated in more detail in Figure 4. The profiles of the seed genes are displayed separately (S), but these genes can be traced back to the intersection of the biclusters. It is important to note that the profiles of the ribosome specific genes **G2** do not line up with the profiles of the cell-cycle genes **G1** in the cell-cycle specific condition set **C1**. Additionally, one can verify that gene-condition combinations **G3**-**C1**, **G3**-**C2** and **G3**-**C3** in the background do not correlate well with the profiles in the biclusters. Condition set **C4** contains those conditions for which the expression of the bicluster genes is not sufficiently coherent or the (average) pattern of the seed genes was too dissimilar from the expression values in **G1** and **G2**. If for example the seeds are underexpressed while the ribosomal genes are overexpressed, the corresponding condition is not included. Because biclustering is robust against noise in the seeds (Figure 2), the algorithm finds the correct gene sets as long as the remaining pattern is significant (this is the case in this example). There is always a trade-off between 'following' the seed and being able to 'correct' the seed. The more informative the priors are, the more our method sticks to the (mean) seed profile.

5.5 Perspectives and future work

© Oxford University Press 2007

The proposed probabilistic framework is flexible and can be extended to a data integration context, by using an appropriate statistical model for each data source. This is a challenge we are currently pursuing.

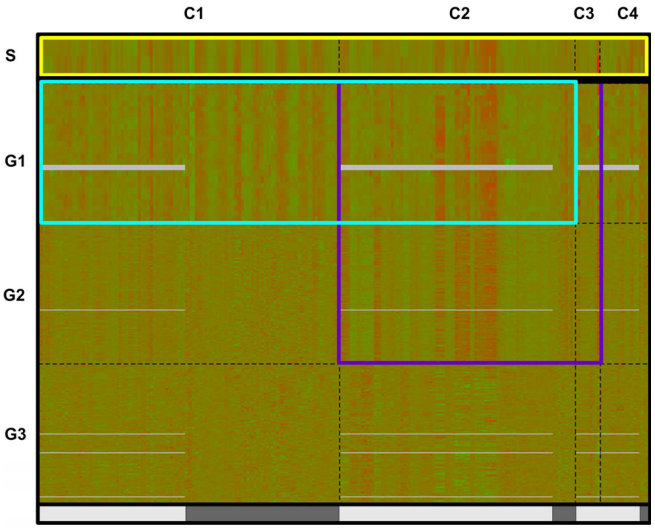


Figure 4: Profiles of bicluster **A** (cyan) and **B** (purple) of Figure 3 (with default gene and condition score thresholds of 0.5), illustrating how the growing bicluster evolves from the cell cycle bicluster to the ribosomal bicluster. The profiles of the seed genes (**S**) are framed in yellow. As expected, the seed genes can be traced back to the intersection (**G1**) of the gene sets belonging to both biclusters. The bar on the bottom indicates which conditions are from the Gasch (light gray) and Spellman (dark gray) data set. For clarity, all gene sets are stretched to similar sizes (**S**: 2 genes, **G1**: 24 genes, **G2**: 1617 genes, **G3**: 4503 genes; **C1**: 123 conditions, **C2**: 96 conditions, **C3**: 10 conditions, **C4**: 21 conditions). Gray indicates missing values in the expression array.

ACKNOWLEDGEMENTS

BDM is a full professor at the Katholieke Universiteit Leuven, Belgium. YM is a professor at the Katholieke Universiteit Leuven, Belgium. TD is research assistant of the Fund for Scientific Research – Flanders (FWO–Vlaanderen). This work is partially supported by: **1**. Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, several PhD/postdoc & fellow grants. **2**. Flemish Government: a. FWO: PhD/postdoc grants, projects G.0407.02, G.0413.03, G.0388.03, G.0229.03, G.0241.04, G.0499.04, G.0232.05, G.0318.05, G.0553.06, G.0302.07, research communities (ICCoS, ANMMM, MLDM); b. IWT: PhD Grants, GBOU-McKnow-E, GBOU-SQUAD, GBOU-ANA, TAD-BioScope-IT, Silicos; SBO-BioFrame **3**. Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011). **4**. EU-RTD: ERNSI; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain.

Conflict of Interest: none declared.

REFERENCES

Agrawal,R. and Imielenski,T. (1993) Mining association rules between sets of items in large databases. In Buneman,P. and Jajodia,S. (eds), *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington D.C.* ACM Press, New York, pp. 207-216.

Ashburner,M. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25-29.

Ben-Dor,A. et al. (2003) Discovering local structure in gene expression data: the order-preserving submatrix problem. *J. Comput. Biol.*, **10**, 373-384.

Bergmann,S. et al. (2003) Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E. Stat. Nonlin. Soft. Matter Phys.*, **67**, 031902.

Bernard,A. and Hartemink,A.J. (2005) Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. *Pac. Symp. Biocomput.*, 459-470.

Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 93-103.

Dempster,A.P. et al. (1977) Maximum likelihood for incomplete data via the EM algorithm. *Journal of Royal Statistical Society*, **39**, 1-38.

Friedman,N. (2004) Inferring Cellular Networks Using Probabilistic Graphical Models. *Science*, **303**, 799-805.

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

Gelman,A.B. et al. (2004) *Bayesian data analysis*. Chapman & Hall/CRC.

Gentleman,R.C. et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Gevaert,O. et al. (2006) Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression. *Hum. Reprod.*, **21**, 1824-1831.

Hubbard,T.J. et al. (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610-D617.

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

Lemmens,K. et al. (2006) Inferring transcriptional modules from ChIP-chip, motif and microarray data. *Genome Biol.*, **7**, R37.

Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM. Trans. Comput. Biol. Bioinform.*, **1**, 24-45.

Murali,T.M. and Kasif,S. (2003) Extracting conserved gene expression motifs from gene expression data. *Pac. Symp. Biocomput.*, 77-88.

Prelic,A. et al. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics.*, **22**, 1122-1129.

R Development Core Team. (2006) *R: A Language and Environment for Statistical Computing*, Vienna, Austria.

Sheng,Q. et al. (2003) Biclustering microarray data by Gibbs sampling. *Bioinformatics.*, **19** (Suppl 2), II196-II205.

Sokal,R.R. and Rohlf,F.J. (1995) *Biometry: the principles and practice of statistics in biological research*. W.H. Freeman and Co, New York.

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

Tanay,A. et al. (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18** (Suppl 1), S136-S144.

Van den Bulcke,T. et al. (2006) Inferring transcriptional networks by mining 'omics' data. *Current Bioinformatics*, **1**, 301-313.

Wu,C.J. and Kasif,S. (2005) GEMS: a web server for biclustering analysis of expression data. *Nucleic Acids Res.*, **33**, W596-W599.