Bioinformatics Advance Access published August 8, 2007

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

Associate Editor: Prof. John Quackenbush

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 introduce QDB, 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 deals with missing values naturally, performs well on artificial data from a recent biclustering benchmark study and has a number of conceptual advantages when compared to existing approaches for focused module search.

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

The availability of large microarray compendia has brought along many challenges for biological data mining. Several supervised and unsupervised methods have been developed for the analysis of such datasets. In particular, probabilistic models have become a popular choice for modeling high throughput genomic data because they allow natural handling of high noise levels (Friedman, 2004). The contribution in the current paper is mainly based on three observations:

- Existing (bi)clustering methods for microarray data analysis often do not answer the specific questions of interest to a biologist. This lack of sharpness has prevented them from surpassing a rather vague exploratory role. Often, biologists have at hand a specific gene or set of genes (seed genes) which they know or expect to be related to some common biological pathway or function. Based on available high throughput data, they may want to recruit additional genes that are involved in that function. In particular, this problem formulation entails 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 may be linked to the same disease?"
- 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 (see Madeira and Oliveira (2004) for a survey), has become a profitable extension to classical cluster analysis. In other words, the fact that some genes are only tightly coexpressed in a subset of experimental conditions (for these experimental conditions, their regulatory program significantly overlaps) should be taken into account. Classical clustering cannot always recover such sets of genes if the patterns are obscured by a large set of irrelevant conditions (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).
- Genes are often involved in several pathways and functions, giving rise to the notion of overlapping transcriptional modules. Both small but highly homogeneous modules (high 'resolution') and larger but more heterogeneous modules (low 'resolution') can be interesting.

The above observations inspired us to develop a Bayesian probabilistic framework for query-driven module discovery in microarray data. Bayesian models have shown promise in providing answers to specific questions or queries, by transforming the knowledge of biologists into prior probability distributions in the model (see, for example, Gevaert *et al.* (2006) and Bernard and Hartemink (2005)). In particular, we focus on the question: "which genes are (functionally) related to the seed genes and which features (conditions) are relevant for this biological function?" We refer to such a set of genes and its relevant conditions with the terms 'bicluster' or 'module'. A resolution sweep approach was designed to resolve the resolution issue and identify overlapping modules that correspond to multiple pathways the query gene may be involved in (multiple regulation).

Only few existing algorithms, such as the (Iterative) Signature Algorithm (Bergmann *et al.*, 2003), Gene Expression Mining Server (Wu and Kasif, 2005) and Gene Recommender (Owen *et al.*, 2003) allow similar directed searches. In the remainder of the paper, we demonstrate the conceptual advantages and efficacy of our flexible QDB (*Query-Driven Biclustering*) modeling framework over these existing approaches. We propose a Condi-

[*]To whom correspondence should be addressed.

tional Maximization approach for model estimation (Gelman *et al.*, 2004) and explain how intuitive choices for the prior distributions lead to a resolution sweep approach. The method is evaluated on a series of artificial data sets. Search strategy and performance are compared with those of the Iterative Signature Algorithm and Gene Recommender. Finally, we discuss results obtained on the combined Gasch *et al.* (2000) and Spellman *et al.* (1998) yeast microarray gene expression data sets.

2 METHODS

2.1 Artificial data

The artificial data were taken from the supplementary website of a recent biclustering benchmark study (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 with standard deviation up to 0.25 (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* (ranging from 0 to 8) 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 (such as discretization) was performed. We did not apply the output filtering procedure described in Prelic *et al.* (2006) to remove heavily overlapping biclusters or limit the number of modules in the output.

In all experiments, the seed consisted of genes correctly belonging to one of 10 artificial modules. We repeated the biclustering process 10 times, each time with randomly selected seed genes 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 Seeds 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.

2.3 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, such as discretization, was carried out.

2.4 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). All known GO-BP labels from Ensembl (Hubbard *et al.*, 2007) were propagated towards the root of the hierarchy. A Benjamini-Hochberg method was used to correct for multiple testing (Storey and Tibshirani, 2003).

3 MODEL AND ALGORITHM

Our general goal is the identification of clusters of genes with similar expression profiles (coordinated changes) in a significant subset of measured experimental conditions (i.e. constant column biclusters in the terminology of Madeira and Oliveira (2004)). By exploiting knowledge contained in a given set of seed genes, we limit the search space through the assumption that the biclusters of interest are those that represent patterns similar to the seed gene pattern (note that this eliminates the need for a masking approach). Because a gene might belong to more than one pathway, we implement a resolution sweep approach to explore a continuum ranging from small but highly homogeneous modules to larger but more heterogeneous modules. Modules at different 'resolutions' might emphasize different aspects of the cellular network. A statistical criterion is used for automatic identification the resolutions of interest.

In the remainder of this section, we first introduce the general modeling framework and briefly discuss a strategy for model estimation. Subsequently, we introduce the query via prior distributions and conclude that a resolution sweep approach is appropriate for the query-driven biclustering problem if the most interesting resolutions are a priori unknown.

3.1 General modeling framework

The core of the probabilistic framework resembles that of (Sheng *et al.*, 2003), the main ingredients being column-wise probability distributions and hidden labels (**g**) for the genes and (**c**) for the conditions to indicate bicluster membership. Assume each column *j* of the (n x m) expression data matrix **X** represents an experimental condition and each row *i* represents a gene. Expression values x_{ij} for which both the corresponding gene and condition are assigned to the bicluster (g_i = 1 and c_j = 1) are then modeled by the bicluster distribution (superscript 'bcl') of the corresponding condition. All other expression values are modeled by the background distribution (superscript 'bgd') of their corresponding condition. The use of condition-wise background distributions allows compensating for between-array differences in expression level variance.

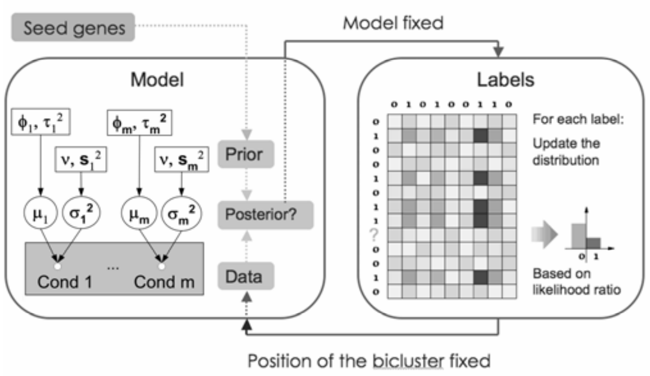

Figure 1: Conceptual scheme of the Bayesian framework for biclustering. On the left, column-wise (condition-wise) Gaussian distributions for the bicluster and background data are represented. Mean and variance parameters are represented by circles and hyperparameters by rectangles. In an iterative procedure, these models are re-estimated while gene and condition labels are assumed fixed (full conditionals for model parameters). On the right, we illustrate how a binary gene label is re-estimated while the models and the other labels are assumed fixed (full conditionals for labels).

Of course, we do not know in advance which genes, conditions and expression values x_{ij} belong to the bicluster; the model therefore depends on hidden gene (**g**) and condition labels (**c**). Figure 1 shows a conceptual scheme of the framework.

For the column-wise (condition-wise) statistical probability distributions we use Gaussian distributions with parameters $\theta_j = (\mu_j, \sigma_j)$ and conjugate Normal - Inverse χ^2 priors. The (full conditional) label probabilities are given by Bernoulli distributions with Beta priors.

3.1.1 Prior distributions

One of the strengths of the Bayesian probabilistic framework is the possibility of using well-chosen prior distributions on the model parameters. We utilize conjugate Normal - Inverse χ^2 priors on the column-wise Gaussian probability distributions (Gelman *et al.*, 2004):

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

If x_{ij} indicates the expression level of gene i in experimental condition j, the corresponding distributions for bicluster and background can then be written as

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

The parameterization of the above formulas is justified by the interpretation of the corresponding full conditional distributions in section 3.1.2 (where parameters κ and ν are equivalent to a number of prior observations). In addition to the prior distributions on the model parameters, Beta priors $\mathrm{B}(\xi_{g1}, \xi_{g0})$ and $\mathrm{B}(\xi_{c1}, \xi_{c0})$ on the parameters of the Bernoulli label distributions can be used to specify a prior believe that a gene or condition belongs to the bicluster (bicluster size).

We postpone the discussion of parameter choices for the priors to section 3.3.

3.1.2 Full conditional distributions

As illustrated in Supplementary File 1, the full conditional distributions for the gene labels are Bernoulli distributions:

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

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

In this expression, ξ_{g1} and ξ_{g0} are parameters of the Beta prior distribution $\mathrm{B}(\xi_{g1}, \xi_{g0})$ on the probability that a gene belongs to the bicluster, n is 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_{g0}, ξ_{g1}, ξ_{c0}, ξ_{c1}) 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 prior knowledge on the bicluster size. 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** (see Supplementary File 1).

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

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

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

and

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

$$\begin{cases} \gamma_j^{\mathrm{bcl}} = \dfrac{\kappa^{\mathrm{bcl}}\varphi_j^{\mathrm{bcl}} + \|\mathbf{g}\|_1 \overline{\mu}_j^{\mathrm{bcl}}}{\kappa^{\mathrm{bcl}} + \|\mathbf{g}\|_1} \\ (\lambda_j^{\mathrm{bcl}})^2 = \dfrac{(\sigma_j^{\mathrm{bcl}})^2}{\kappa^{\mathrm{bcl}} + \|\mathbf{g}\|_1}. \end{cases}$$

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 $\overline{\mu}_j^{\mathrm{bcl}}$ (variance) and the prior mean φ_j^{bcl} (variance). For brevity, we omitted the formulas for the background model parameters, which are similar. Details on the derivations can be found in Supplementary File 1.

3.1.3 Joint posterior distribution

Given the data and a particular choice for the prior distributions, the joint posterior distribution $p(\boldsymbol{\theta}, \mathbf{g}, \mathbf{c} \mid \mathbf{X}, \boldsymbol{\psi})$ indicates the statistically most interesting simultaneous assignments for the labels and the model parameters. Unfortunately, it is not possible to use this joint posterior distribution directly because we are unable to describe it analytically. We next discuss a strategy to detect its local maxima using information on the corresponding full conditional probability distributions only.

3.2 Algorithm

Conditional Maximization or CM (Gelman *et al.*, 2004) consists of alternatively maximizing a set of full conditional distributions. These maximization steps are repeated until the procedure converges to a local mode of the corresponding joint probability distribution. Figure 1 illustrates the alternating procedure; the full conditionals and joint posterior distributions were introduced in sections 3.1.2 and 3.1.3 respectively. In the query-driven context under study, convergence to local modes in the posterior landscape will often be sufficient because the query is introduced through strong priors on the model parameters (see section 3.3). Strong priors act as a powerful zoom lens to magnify interesting regions in the likelihood landscape. Therefore, they tend to give rise to rather simple posterior distributions, even when the corresponding likelihood landscape is complex and contains many modes. Furthermore, the knowledge represented by the seed genes can be used for clever initialization (Supplementary File 1).

3.3 Introducing the query

The discussion in Section 3.1 applies to all possible instantiations of models in the general framework. Before tying up any prior parameters based on the knowledge of the seed genes, 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 priors, which we deemed intuitive. One should keep in mind that alternative ways to introduce the prior knowledge exist.

3.3.1 Different kinds of priors

Strictly speaking, 4 groups of parameters specify the priors:

(1) The parameters for the priors on the means and variances of the bicluster models

(2) The parameters for the priors on the means and variances of the background models

(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 a condition belongs to the bicluster

Although these parameters have a clear-cut statistical interpretation, it is not desirable to choose all of them manually. Therefore, we decided to fix most of them by default:

- Assume that the prior probability of a gene belonging to the bicluster is unknown. The corresponding Beta prior can then be fixed to B(1,1) (noninformative, uniform distribution), eliminating (3).
- For the Beta prior $B(\xi_{c1}, \xi_{c0})$ on the condition labels, ξ_{c1} equals twice the total number of conditions in all experiments while ξ_{c0} was fixed and equal to one in all experiments (all artificial and real data sets). This setting forces (at least) those conditions into the background for which the bicluster genes do not have a significantly better likelihood under the bicluster distribution. The exact setting of this prior does not have much influence as long as $\xi_{c0} > \xi_{c1}$ (data not shown).
- To sufficiently restrict the freedom of the bicluster (preventing it from drifting too far away from the seed profile), 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, tying up additional parameters in (1).
- For the priors on the means and variances of the background models, we use $\varphi_j^{bgd} = \bar{\mu}_j$ and $s_j^{bgd} = \bar{\sigma}_j$ with a number of pseudocounts κ^{bgd} and ν^{bgd} equal to the total number of genes in the data set. The exact number of pseudocounts has little influence for reasons explained on the Supplementary website.

When the other prior distributions are fixed as described above, the remaining bicluster variance priors (one for each condition) provide sufficient flexibility to accommodate various query-driven biclustering strategies. As explained in section 3.1.1, these priors are (scaled) inverse χ^2 distributions with two kinds of parameters: $(s_j^{bcl})^2$ 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 procedure, 90 prior observations would mean that the resulting variance of the bicluster is determined for 90% by the prior variance and for 10% by the variance of the 10 genes currently in the bicluster.

3.3.2 Noninformative variance priors

In most artificial data scenarios, results are 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) perform well here, since the patterns are very strong in most cases and this strength is (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 $(s_j^{bcl})^2$, for instance equal to background variance) and still detect most of the modules (data not shown).

3.3.3 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. Therefore, stronger priors are needed to extract interesting but 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 large 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 can be exceeded by weaker correlation between larger numbers of genes over a smaller subset of the conditions. It may happen that the profiles of the seed genes match better with the expression pattern of a dominant bicluster than with the background pattern over a sufficiently large number of conditions, causing the algorithm to get stuck in the corresponding mode of the posterior distribution. In such cases, it seems preferable to be able to explicitly zoom in on more appropriate modules by using informative variance priors as a control handle. 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 informative settings for the prior on the variance. We propose a resolution sweep approach in which the prior variance is slowly (and linearly) increased while the algorithm is running. In fact, this means that the starting point at each prior setting is equal to the posterior mode that was found with a slightly smaller value of the variance prior, a sensible initialization. In other words, we start close to the seed (Supplementary File 1) and stay in a mode of interest at any time. Because the algorithm remains near convergence, only a few iterations are needed for each prior variance setting.

Varying prior variance corresponds to traveling through the modular structure of the data in the neighborhood of the seed (see Supplementary File 1 for examples and an intuitive comment on the notion of resolution). In both the real and artificial data sets, linear increases of the prior parameter result in discrete steps for the observed module sizes, illustrating crisp transitions between modules at different resolutions. In Section 5.2.1, we report on an example that evolves from very specific cell-cycle related functions over less specific ones to ribosome biogenesis related functions (when the prior variance is large). The corresponding figures (and many more examples on the supplementary website) 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 the resolution sweep approach. In artificial data, we linearly increased the variance prior parameter $(s_j^{bcl})^2$ for each condition from 0 to the corresponding background variance over 100 iterations. In the yeast data sets, we used the same strategy but over 2000 iterations. The variance priors were chosen to be very informative by setting the number of pseudocounts ν^{bcl} equal to the total number of genes. To automatically detect the resolutions of interest, we identified the local maxima in the *Akaike Information Criterion* (AIC) (Akaike, 1974) for model selection on the resolution sweep path (see

Supplementary File 1 for a rationale on this criterion for automatic module detection). Since the pattern search is centered on the mean seed gene pattern, it is not surprising that the seed itself corresponds to one of those local optima in many cases. Therefore we consider the seed to be a trivial module and exclude it from the output. As explained in the Systems and Methods section, no additional postprocessing was carried out.

Note that we do not only report the module with the maximal AIC score for each seed. Indeed, modules that are statistically most relevant are not always most interesting. In the yeast data set for example, dominant ribosome biclusters containing many genes often have the highest score. Nevertheless, it is worth noting that in the artificial data scenarios under study the best scoring module almost always corresponded to the 'correct' module.

4 IMPLEMENTATION

All algorithms were implemented in the R language and environment for statistical computing (R Development Core Team, 2006). An implementation of the Gene Recommender algorithm, the GO Biological Process functional annotations and the multiple testing correction of the p values were obtained using packages from the Bioconductor repository (Gentleman *et al.*, 2004).

5 RESULTS AND DISCUSSION

5.1 Artificial expression data

In order to make a fair comparison and avoid any bias in creating our own data sets, we systematically evaluated our QDB algorithm on two artificial data scenarios (S1 and S2) from a recent biclustering benchmark paper (Prelic *et al.*, 2006), containing noiseless overlapping modules (A) and noisy non-overlapping modules (B). More details on the setup and the definition of the performance measures can be found in the Systems and Methods section. To make the benchmark more informative, we included results of the Iterative Signature Algorithm (ISA) and Gene Recommender (GR) using the same seeds. A short description of the ISA and GR algorithms can be found in Supplementary File 1. Before discussing numerical results, we highlight some fundamental conceptual differences between QDB and the ISA and GR algorithms.

5.1.1 *Conceptual differences - Iterative Signature Algorithm*

The parameter variation approach suggested by the authors of the Iterative Signature Algorithm (Ihmels et al., 2004) is similar in spirit to the resolution sweep approach presented here. However, our approach is different in some important ways:

- ISA is a clever algorithm rather than a solid probabilistic modeling framework. Modules are defined as fixed points of this algorithm without referring to an underlying mathematical model or explicit cost function. This makes it difficult to extend ISA to include other data sources, where the algorithm may not be directly applicable. The QDB framework is flexible in allowing the specification of other distribution types or search strategies.
- The probabilistic interpretation of QDB allows automatic selection of interesting modules on the resolution sweep path, via local optima in the AIC score. In contrast, ISA does not have a natural notion of statistical scores for the reported modules.
- The search strategy of ISA is based on significant average overexpression or underexpression of the bicluster genes in the bicluster conditions whereas the QDB search strategy is based on significant differences in expression between the bicluster and the background.
- In contrast to ISA, QDB deals with missing values naturally (Supplementary File 1).
- ISA is query-based in the initialization only. Strictly speaking, there is no guarantee that the module does not drift away from this query point, due to the presence of more dominant modules nearby. Some experiments on real data indeed revealed modules that did not contain any of the query genes at any resolution (as specified by threshold parameter t_G) of interest. In contrast, by explicitly controlling the statistics (for example mean and variance) of the bicluster to a degree specified by the prior strength, it is straightforward to prevent QDB from reporting biclusters that are too remote from the query.

5.1.2 *Conceptual differences - Gene Recommender*

Gene Recommender (Owen *et al.*, 2003) is designed to prioritize genes rather than to detect transcriptional modules. The output of the core algorithm is an ordered gene list. In order to convert this output into a module format, an appropriate cutoff has to be specified. The default threshold of the Gene Recommender software corresponds to 50% recall, but on small seed sets (for example two genes) this yields trivial results, mostly modules containing one (seed) gene only. Therefore, we decided to show GR results with optimal thresholds: for every seed we selected those thresholds that correspond to optimal module recovery and bicluster relevance scores, resulting in an upper bound on GR's performance. It is important to keep this in mind when interpreting the numerical results in the next section.

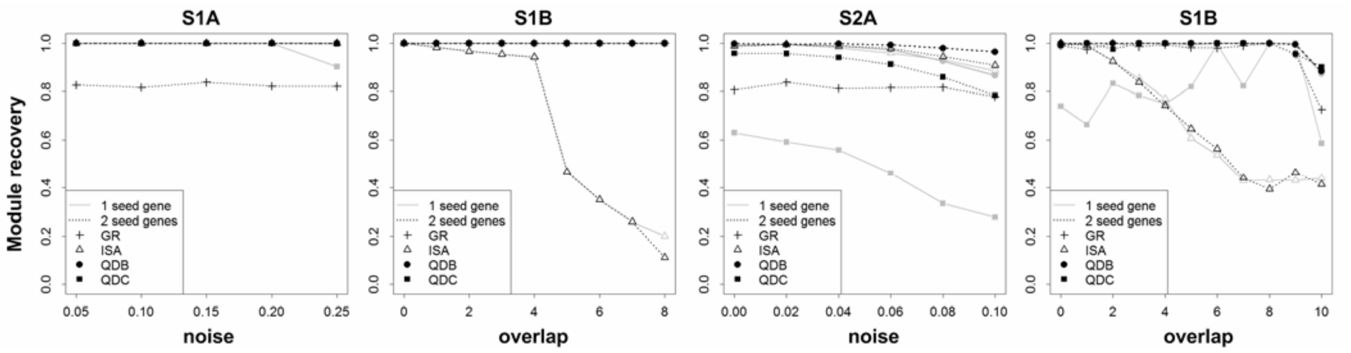

Figure 2: Evolution of module recovery scores as a function of noise (indicated by the letter A) and overlap (letter B), in two artificial data scenarios (S1 and S2), taken from (Prelic *et al.*, 2006). Results obtained with seeds of one gene are shown with gray solid lines, results with seeds of two genes in black dotted lines. QDC refers to an additional clustering variant of our own query-driven Bayesian framework (see main text).

Three more advantages of QDB over GR are worth mentioning in a module discovery context:

- In contrast to QDB, GR is unable to deal with queries of one gene only.
- The GR output does not contain as much information as the QDB output on the structure of the modules over different resolutions. Although a variation of the cutoff parameter on the output list corresponds to some notion of resolution, the condition content remains the same and no genes are allowed to drop out upon increasing the threshold. Therefore, the notion of overlapping modules at different resolutions does not exist as such in the Gene Recommender system.
- GR first selects appropriate conditions and then uses the correlation of candidate genes in the selected conditions to obtain a ranking of the genes with respect to the seed. While QDB simultaneously selects genes and conditions, this is not the case for GR.

5.1.3 Simulation results

Figure 2 shows module recovery scores for small seeds of one and two genes. Due to a lack of space, we moved the corresponding bicluster relevance plots to our supplementary material (Supplementary Figure 1).

In general, query-based algorithms perform very well on these data compared to various global biclustering methods (Prelic *et al.*, 2006). This illustrates that the use of a query can facilitate the search, even though the setup uses small data sets with modules of nearly equal strength and therefore does not fully exploit the advantages of the query-based approaches. For every scenario, scores are comparable with the best scores in Prelic *et al.* (2006).

To make the comparison more informative, we included a query-based clustering (in contrast to biclustering) variant of our algorithm, by simply removing the condition selection. As expected, the biclustering variant outperforms the clustering variant, especially in noisy scenarios. This demonstrates that removing irrelevant conditions can substantially improve the results.

The performance of QDB and ISA is nearly equal in the noise scenarios. In the overlap scenarios, QDB outperforms ISA, the main reason being the multi-resolution aspect and the automatic detection of the most relevant resolutions in QDB. Moreover, QDB modules with the maximum AIC score were 'correct' modules in almost all scenarios, supporting the use of the AIC scores as a measure of statistical relevance. ISA is unable to report the most relevant modules in the overlap scenarios because they do not correspond to the resolution represented by the (default) gene and condition resolution settings. Even when used in a parameter variation setting (as in Ihmels *et al.* (2004)), ISA does not have a natural notion of statistical scores for the reported modules.

GR possibly performs well on overlap scenarios but is outperformed by both ISA and QDB in noisy scenarios. Recall that the plots show upper bounds on GR's performance only.

5.2 Yeast expression data

Although artificial data are helpful to gain understanding in properties of algorithms, they remain an approximation of biological reality. Therefore, the performance of our approach was further examined by applying the QDB resolution sweep algorithm to a concatenation of two well-known yeast expression compendia (Gasch *et al.* (2000) and 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).

The results of our analyses of over 100 sets of seed genes can be found on our supplementary website. In most cases, we were able to 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. For more details, we refer to the supplementary website.

5.2.1 A cell cycle bicluster example

Figure 3 and Figure 4 show an example of overlapping modules that were detected using one of the seeds. The seed consisted of two genes (IRC8 and CDC5) 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). When the prior variance is increased, the number of selected conditions decreases while the number of genes starts to increase. The first selected module (**A**) contains genes which are involved in DNA-dependent DNA replication. However, it does not have a significant functional overrepresentation after correcting for multiple testing. When we further increase the variance, the algorithm picks up the signal from stronger mitotic cell cycle (p_{corr} = 1.6e-3) and cell division (p_{corr} = 4.1e-7) modules. Around iteration 1300, there is an abrupt transition to an overlapping ribosomal module (p_{corr} < 1e-16). The latter change (from module **F** to **G** 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 cell cycle conditions (Spellman *et al.*, 1998) are lost, in agreement with the change in function.

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 **G3** do not line up with the profiles of the cell-cycle genes **G1-G2** in the cell-cycle specific condition set **C1**. Additionally, one can verify that gene-condition combinations **G4-C1**, **G4-C2** and **G4-C3** in the background do not correlate well with the profiles in the biclusters. The remaining conditions (**C4**) do not belong to either bicluster because 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-G2-G3**. Note that there is always a trade-off between following the seed and allowing deviations from the seed pattern based on evidence in the data. The more informative the priors are, the more our method sticks to the (mean) seed profile.

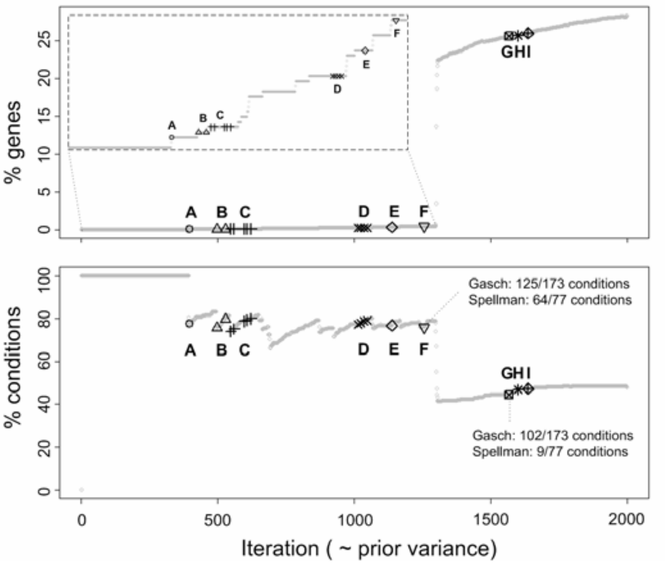

Figure 3: Evolution of bicluster S9 (based on Spellman *et al.* (1998) seed number 9) with increasing prior variance from left to right. Letters indicate the selected modules at various resolutions. Modules with the same gene content were grouped in one letter symbol. Functionally enriched GO Biological processes for the selected modules range from mitotic cell cycle (**D**: p = 1.46e-6, p_{corr} = 1.6e-3) over cell division and cytokinesis (**E-F**: p = 1.99e-10, p_{corr} = 4.1e-7) to ribosome biogenesis and assembly (**G-H-I**: p_{corr} < 1e-16) as the prior variance increases.

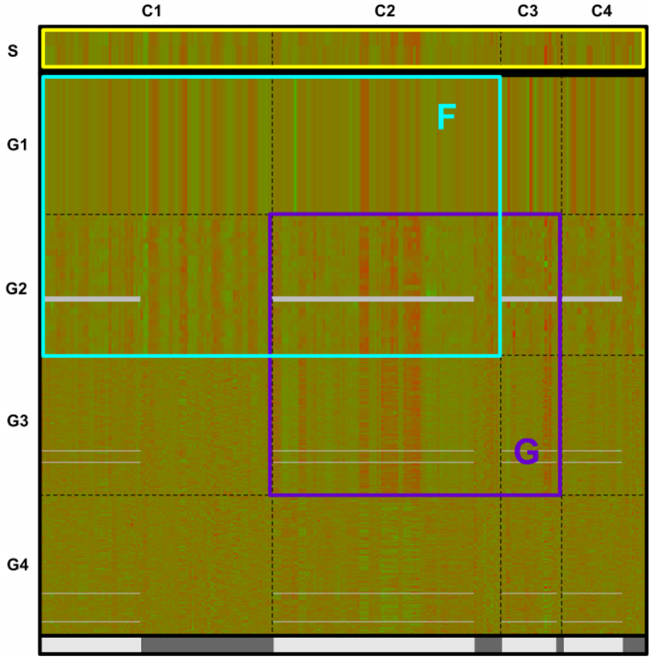

Figure 4: Profiles of bicluster F (cyan) and G (purple) of Figure 3, illustrating the transition from a cell cycle bicluster into a ribosomal bicluster. The profiles of the seed genes (**S**) are framed in yellow. As expected, the seed genes can be traced back to the intersection (**G2**) of the gene sets belonging to both biclusters. The bar on the bottom indicates which conditions are from the Gasch (light gray) and Spellman (dark gray) data sets. For clarity, all gene sets are stretched to similar sizes (**S**: 2 genes, **G1**: 1 gene, **G2**: 24 genes, **G3**: 1550 genes; **G4**: 4569 genes; **C1**: 100 conditions, **C2**: 90 conditions, **C3**: 21 conditions, **C4**: 39 conditions). Missing values in the expression array are indicated in gray.

5.3 Perspectives and future work

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

ACKNOWLEDGEMENTS

TD is research assistant of the Fund for Scientific Research – Flanders (FWO–Vlaanderen). BDM is a full professor at the Katholieke Universiteit Leuven, Belgium. KM is a professor at the Katholieke Universiteit Leuven, Belgium. YM is a professor at the Katholieke Universiteit Leuven, Belgium. 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

Akaike,H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716-723.

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

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.

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. (2004) Defining transcription modules using large-scale gene expression data. *Bioinformatics.*, **20**, 1993-2003.

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.

Owen,A.B. et al. (2003) A gene recommender algorithm to identify coexpressed genes in C. elegans. *Genome Res.*, **13**, 1828-1837.

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**, 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.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A*, **100**, 9440-9445.

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.