BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 14 2006, pages 1737–1744
doi:10.1093/bioinformatics/btl184

Gene expression

Context-specific infinite mixtures for clustering gene expression profiles across diverse microarray dataset

X. Liu[1,2], S. Sivaganesan[3], K. Y. Yeung[4], J. Guo[1], R. E. Bumgarner[4] and Mario Medvedovic[1,2,*]

[1]Department of Environmental Health, University of Cincinnati, 3223 Eden Avenue ML 56, Cincinnati OH 45267, USA, [2]Division of Biomedical Informatics, Cincinnati Children's Hospital Research Foundation, Cincinnati, OH 45229, USA, [3]Mathematical Sciences Department, University of Cincinnati, Cincinnati, OH 45221, USA and [4]Department of Microbiology, University of Washington, Seattle, WA 98195, USA

Received on March 6, 2006; revised on April 18, 2006; accepted on May 8, 2006
Advance Access publication May 18, 2006
Associate Editor: Golan Yona

ABSTRACT

Motivation: Identifying groups of co-regulated genes by monitoring their expression over various experimental conditions is complicated by the fact that such co-regulation is condition-specific. Ignoring the context-specific nature of co-regulation significantly reduces the ability of clustering procedures to detect co-expressed genes due to additional 'noise' introduced by non-informative measurements.

Results: We have developed a novel Bayesian hierarchical model and corresponding computational algorithms for clustering gene expression profiles across diverse experimental conditions and studies that accounts for context-specificity of gene expression patterns. The model is based on the Bayesian infinite mixtures framework and does not require a priori specification of the number of clusters. We demonstrate that explicit modeling of context-specificity results in increased accuracy of the cluster analysis by examining the specificity and sensitivity of clusters in microarray data. We also demonstrate that probabilities of co-expression derived from the posterior distribution of clusterings are valid estimates of statistical significance of created clusters.

Availability: The open-source package gimm is available at http://eh3.uc.edu/gimm

Contact: Mario.Medvedovic@uc.edu

Supplementary information: http://eh3.uc.edu/gimm/csimm

1 INTRODUCTION

Identifying and interpreting gene expression patterns and characterizing groups of co-expressed genes defining these patterns through cluster analysis has been a productive approach to learning from DNA microarray data. The results of such analyses have served to dissect regulatory mechanisms underlying co-expression, identify pathways involved in biological processes and annotate gene function. The quality of these results and conclusions are directly dependent on the quality of the clustering procedure used in the analysis. Since the advent of the microarray technology, virtually all traditional clustering approaches have been applied in this context and numerous new approaches have been developed (Yeung and Bumgarner, 2004). To identify subsets of co-expressed genes, most clustering procedures depend on either a visual identification clusters from patterns in a color-coded display (such as hierarchical clustering) or on the correct specification of the number of patterns present in data prior to the analysis (*k*-means and Self Organizing Maps). The most commonly used clustering procedures are ad hoc by nature and incapable of separating statistically significant clusters from artifacts of random fluctuations in the data. On the other hand, clustering approaches based on the statistical modeling of the data often require the number of clusters to be specified in advance (Barash and Friedman, 2002; McLachlan *et al*., 2002; Segal *et al*., 2003). When the 'correct' number of clusters is estimated from the data, traditional methods fail to account for this significant source of variability in assessing the statistical significance of detected patterns (Medvedovic and Guo, 2004).

Assessing the function of a gene product is a multidimensional endeavor whereby one may ascertain a number of properties including structure, the low-level function of a protein (i.e. kinase, protease, etc.) and a higher-level function describing the biological processes in which the protein participates. Identification of groups of co-expressed genes across diverse microarray datasets is a very promising strategy for assessing higher-level function of gene products. Such analysis is complicated by the fact that co-regulation is often condition-specific and may not extend across all conditions. The problem of context-specificity can be particularly pronounced when combining gene expression profiles across different experiments, tissue types or even different organisms to perform 'meta-cluster analysis' (Segal *et al*., 2004, 2003; Stuart *et al*., 2003). In these situations, measurements of genes' expression under all conditions are not necessarily informative with regards to their co-regulation. Ignoring the local nature of co-regulation, significantly reduces one's ability to detect co-regulated genes owing to the noise introduced by non-informative measurements. Previously proposed solutions to this problem in terms of the context-specific Bayesian networks (Barash and Friedman, 2002) and more general module networks (Segal *et al*., 2003) rely on the specification or estimation of the 'correct' number of patterns. In these respects, they suffer from the same problems related to the estimation of the 'correct' number of patterns in the finite mixture-based clustering.

*To whom correspondence should be addressed.

© The Author 2006. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from http://bioinformatics.oxfordjournals.org/ at Princeton University on October 9, 2014

We developed a context-specific infinite mixture model (CSIMM) to allow clusters of co-expressed genes to be further grouped locally on subsets of experimental conditions that do not contribute any information about their differences. This approach makes use of the Bayesian infinite mixture framework (Medvedovic *et al*., 2004; Medvedovic and Sivaganesan, 2002) to circumvent the issue of identifying the 'correct' number of global and local patterns in the data. Infinite mixtures are one possible parametrization of semi-parametric Bayesian models with Dirichlet process priors (Neal, 2000) and the CSIMM described here can be thought of as a hierarchical Dirichlet process. Infinite mixtures framework facilitates averaging over models with different numbers of patterns, and the posterior distribution of clusterings incorporates uncertainties related to not knowing the 'correct' number of clusters, either globally or locally. Consequently, the resulting posterior probabilities of co-expression offer a reliable assessment of the statistical significance of the groupings. We demonstrate the ability of the procedure to integrate information from diverse microarray experiments through a simulation study and by assessing the performance of algorithms in the context of functional annotation of genes based on their co-expression.

2 METHODS

2.1 Motivation

Our goal is to identify clusters of genes exhibiting similar expression patterns across multiple microarray datasets. Each dataset or 'context' consists of a number of closely related microarray experiments that share a biological relationship and sample a limited range of perturbations to the system under study. For example, one dataset may consist of measurements of gene expression at different time points after heat shock, while another may consists of measurements at different stages of the cell cycle. For the sake of discussion, we will refer to each dataset as a 'context' and the entire collection of datasets as the 'global' dataset.

It is reasonable to assume that different regulatory programs are employed by different biological processes and that specific subsets of regulatory programs are needed to respond to a given type of perturbation. Some regulatory programs will respond to all the perturbations available within the global dataset while others will respond to none, one or a limited number of perturbations. That is, some genes will be co-regulated on a global scale, while others (perhaps most) will be co-regulated on a local scale. We define a global clustering structure on a set of gene expression profiles by saying that two genes belonging to the same global clusters share a common pattern of expression in all of the examined datasets. On the other hand, we define a local clustering structure as groups of genes that share a common expression pattern within a subset (a context) of the data but which do not group together when examined globally.

Figure 1 shows an example of the type of structure we might reasonably expect within gene expression data and the information (groupings) we would like to recover. For clarity, the example provided is overly simplified. It shows only three datasets (contexts), 20 genes, 4 global clusters and 2 local clusters within each context. Expression level is either high (coded light gray) or low (black). Local clusters within each context consist of genes that are either high or low within this context. For example, global Cluster 1 is formed of genes that are high within Context 1 and low within Contexts 2 and 3. On the other hand, Context 1 does not contribute any information about differences between global Clusters 1 and 4. We wish to be able to separate patterns in such global clusters even in the presence of data from many other 'noisy' or non-informative contexts. We construct a Bayesian hierarchical model that describes the probability distribution of the data so that local and global clusters can be identified.

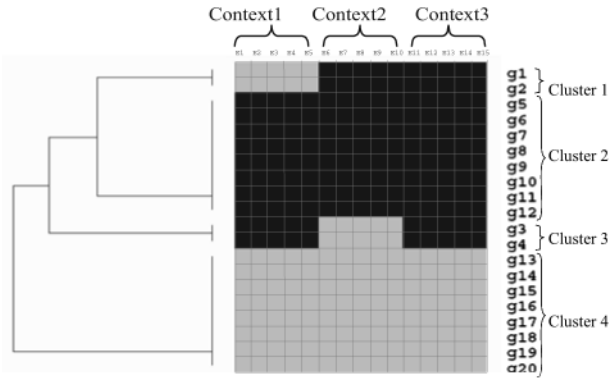

Fig. 1. Simple 'context-specificity' of expression patterns.

2.2 Context-specific infinite mixture model

Suppose that a expression was measured for T genes across M experimental conditions. If x_{ij} is the expression level of the i-th gene for the j-th experimental condition, then $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iM})$ denotes the complete expression profile for the i-th gene. Each gene expression profile can be viewed as being generated by one out of Q different underlying expression patterns. Expression profiles generated by the same pattern form a cluster of similar expression profiles. If c_i is the classification variable indicating the pattern that generates the i-th expression profile ($c_i = q$ means that the i-th expression profile was generated by the q-th pattern), then a clustering is defined by a set of classification variables for all gene expression profiles $\mathbf{C} = (c_1, c_2, \ldots, c_T)$. In our model, the q-th 'pattern' is represented by the mean vector of the M-dimensional Gaussian distribution $\mu_q = (\mu_{q1}, \ldots, \mu_{qM})$. Profiles clustering together (i.e. belonging to the same pattern) are assumed to be a random sample from the same multivariate Gaussian distribution. That is, $c_i = q$ implies that $\mathbf{x}_i \sim N_M(\mu_q, \mathbf{\Sigma}_q)$, where $\mathbf{\Sigma}_q$ is the variance–covariance matrix of the M-dimensional multivariate Gaussian distribution.

Suppose further that each gene profile is partitioned into R sub-profiles. Without loss of generality we can assume that the first r_1 experimental conditions form the first sub-profile, experimental conditions r_1+1 to r_1+r_2 form the second sub-profile and so on. That is $\mathbf{x}_i = (\mathbf{x}_i^1, \ldots, \mathbf{x}_i^R)$ where $\mathbf{x}_i^j = (x_{ir'_j+1}, \ldots, x_{ir'_j+r_j})$ and $r'_j = r_1 + \cdots + r_{j-1}$. The two most extreme cases are when $R = M$ and $R = 1$. The case of $R = 1$ is equivalent to the simple clustering in which the context structure is not defined. The case when $R = M$ represents the situation when each microarray hybridization represents a separate context. The local structure of the co-expression patterns is specified by the Q by R matrix $\mathbf{L} = (L_{qf})$, where $L_{qf} = t$ if global cluster q is placed in local cluster t within context f. Thus, within each context, we create a group of 'locally' indistinguishable global clusters. All gene expression profiles contained in global clusters that are indistinguishable within a context form a local cluster of genes which are co-expressed within this context.

The joint posterior distribution of all parameters in the model, including the global and local clustering variables $\mathbf{C}$ and $\mathbf{L}$, given data are estimated using the Gibbs sampler (Gelfand and Smith, 1990). The clusters of globally and locally co-expressed genes are formed based on the marginal posterior distributions of the classification variables $\mathbf{C}$ and $\mathbf{L}$. Summarizing the sample of ''clusterings'' generated by the Gibbs sampler in mixture models is generally a non-trivial problem. We circumvent this problem by calculating posterior pairwise probabilities (PPPs) of co-expression for genes i and j as the proportion of the samples in which these two genes are clustered together (Medvedovic and Sivaganesan, 2002). We then use these probabilities as the similarity measure to hierarchically cluster gene expression profiles by applying the average linkage principle. The mathematical specification of the model describing the distribution of the data and the specifics of the Gibbs sampler are given in the Appendix. All conditional probability

Downloaded from http://bioinformatics.oxfordjournals.org/ at Princeton University on October 9, 2014

distributions needed to run the Gibbs sampler are given in the Supplementary Materials.

2.3 Implementation

Computational procedures for performing CSIMM-based clustering are implemented in a standalone package *gimm*. The package consists of C++ code, simple Java-based gui and installation scripts, and it is available for both Linux/Unix and Windows platforms. The Windows version is available as a self-installing package. The software generates .cdt and .gtr files defining the hierarchical clustering that can be viewed and analyzed using the treeview program (Eisen *et al*., 1998). The Linux C++ code is designed to exploit the OpenMP parallelization when appropriate compiler is installed. For Linux, we also developed the R package 'wrapper' that facilitates using *gimm* within R. All packages and the source code can be freely downloaded from http://eh3.uc.edu/gimm. We discuss the computational complexity of the algorithm in the Supplementary Material.

3 RESULTS

3.1 Simulation study

This study was designed to compare different clustering procedures based on their ability to correctly separate simulated expression profiles into different clusters in repeated experiments. The problem is treated in the traditional statistical hypothesis-testing framework of assessing the probability that a procedure will correctly conclude that two expression profiles are generated by distinct patterns of expression (i.e. belong to two different clusters) while controlling the probability of falsely concluding that two profiles belong to different clusters when they are actually generated by the same pattern. Unlike traditional statistical hypothesis-testing procedure, we do not supply the labels for profiles that are being compared. We simulated data representing the structure depicted in Figure 1 where the heat map was taken to represent the values of the mean expression profiles in the corresponding cluster. 'Low expression levels' (black) were set to 0 and the 'high expression levels' (gray) were set to 1. For example, in each dataset, profile 'g1' was randomly drawn from the 15-dimensional Gaussian random distribution whose mean vector is equal to 1 in first 5 dimensions ($e1, \ldots, e5$) and 0 in other 10 dimensions ($e6, \ldots, e15$). Data were simulated for different level of noise (σ). The selected range of random noise allowed us to assess the performance of different approaches in easy and progressively more difficult (i.e. noisier) situations. A total of 100 datasets were generated for each noise level.

We are focusing on the ability of a method to separate profiles in Cluster1 from profiles in Cluster2. This is the most difficult aspect of the analysis since Cluster1 has only two profiles and differs from Cluster2 only on 5 out of 15 'experimental conditions'. Methods are tested based on their ability to correctly conclude that profiles in Cluster1 are different from profiles with Cluster2. In a sense we are assessing the power of clustering procedures to conclude that profiles in Cluster1 are different from profiles in Cluster2. If we knew which profiles came from which cluster, we could perform a simple test of hypothesis to decide one way or another. In the unsupervised situation, we do not supply the membership but the goal is still the same. The performance of different clustering procedures was assessed by constructing receiver operating characteristic (ROC) curves that relate the probability of the clustering method to correctly separate profiles from different clusters and the probability of incorrectly separating profiles from the same clusters.

Let X be the posterior probability cut-off for separating profiles in Cluster1 from Cluster2. For a fixed cut-off point X, we consider that the clustering procedure is correctly concluding that a profile from Cluster1 does not belong to Cluster2 if its posterior pairwise probability of co-clustering with any single profile from Cluster2 is less than X. That is, $\max\{p(c_i = c_j \text{ for all profiles } j \text{ from Cluster2}\} < X$, where $p(c_i = c_j)$ denotes the posterior pairwise probability of co-expression for profiles i and j. We consider that the clustering procedure is incorrectly concluding that profile 1 and profile 2 from Cluster1 do not belong in the same cluster if $p(c_1 = c_2) < X$. The true positive rate (TPR) is the proportion of times that a correct decision is made and the false positive rate (FPR) is the proportion of times that an incorrect decision is made. As the cut-off X is increased from 0 to 1, both TPR and the FPR will increase. The area under the curve relating the TPR and FPR as X is increased from 0 to 1 describes the efficiency of a statistical procedure with the random decision-making having an area of 0.5 while the ideal statistical procedure would have an area equal to 1. ROC curves in Figure 2 indicate that context-specific infinite mixtures model significantly outperforms simple mixture model in its ability to separate different patterns of expression while controlling the false positive rates. The difference between the simple infinite mixture and context-specific infinite mixtures is clearly due to the better representation of the underlying patterns offered by the context-specific model. Furthermore, over- and under-fitting the data by specifying too many or too few context has the expected consequences on the clustering results (Supplementary Figure S1). When placing each experiment in separate context (over-fitting), the performance is actually worse than for the simple model. Failing to specify all contexts (two out of three) causes a reduction in the performance of the context-specific model, but it still outperforms the simple model.

Posterior pairwise probabilities are valid measures of statistical significance. In Figure 3 we plotted observed false positive rates against corresponding statistical significance levels from the CSIMM analysis. Given a significance level α, all gene-pairs whose PPP was lower than α were assumed to belong to different clusters. As the empirical false positive rates are always less than α, we conclude that PPPs based on CSIMM are valid measures of statistical significance at all noise level. This is also true for the simple IMM, but not for the finite mixtures model (Supplementary Figure S2). In addition we performed similar analysis on 100 datasets in which all data points were generated from the single probability distribution representing the situations when there is no clustering structure in the data (Random). As it can be seen from the virtually perfect 45° line, PPPs correctly protect against Type I errors when there are no patterns in the data.

3.2 Yeast cell-cycle and sporulation data

Comparing the performance of different clustering procedures on the real-world data is complicated by the lack of a gold-standard (i.e. the 'correct clustering'). We assessed our clustering results by forming functional groupings of genes based on the information available in the KEGG database of biological pathways (Kanehisa *et al*., 2004). The strength of association between such functional clusters and clusters of co-expressed genes formed by a clustering procedure was interpreted as the measure of the precision for a clustering procedure.

Downloaded from http://bioinformatics.oxfordjournals.org/ at Princeton University on October 9, 2014

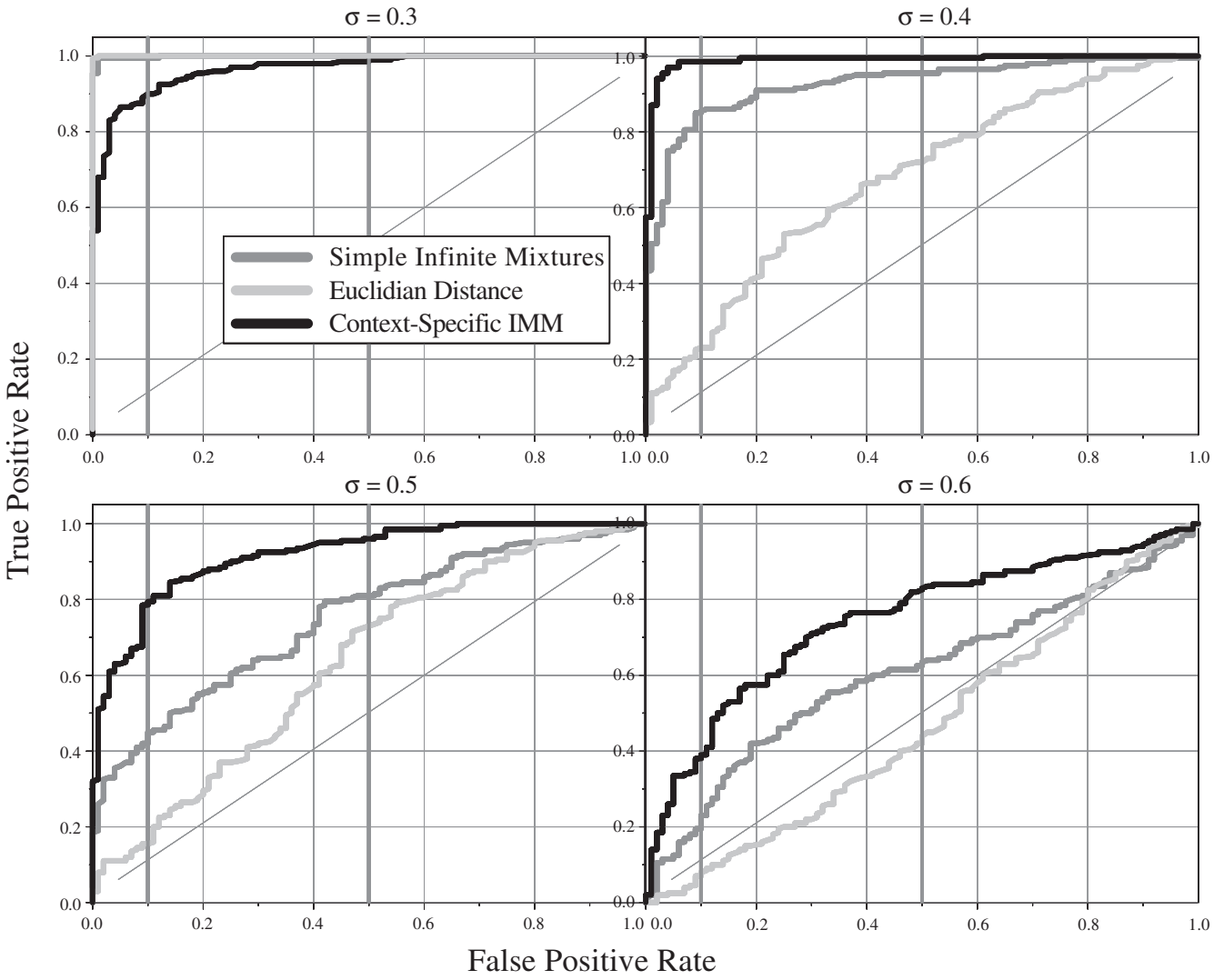

Fig. 2. ROC curves for different clustering approaches.

We constructed the test dataset by combining two microarray experiments assessing two distinct yet related biological processes. The first dataset is the yeast sporulation dataset (Primig *et al*. 2000) consisting of gene expression measurements at 8 and 7 time points throughout the sporulation process for two sporulation-competent yeast strains SK1 and W303, respectively. The second dataset is the cell-cycle (Cho *et al*., 1998) dataset consisting of gene expression measurements at 17 time-points spanning two complete yeast cell cycles. The two datasets were matched by identifying 6044 ORFs represented on both of the two versions of the Affymetrix microarrays used in these experiments. Data was mildly processed by setting any measurement below 1 to 1, log-transforming it and centering each gene's expression profile around zero for the two experiments separately. Genes which never reached the signal of 100 were excluded from the analysis resulting in the total of 5685 genes remaining. A total of 1044 ORFs represented genes associated with at least one KEGG pathway.

Data from the two experiments were clustered separately and jointly using the simple IMM approach, CSIMM and Euclidean distance-based hierarchical clustering (EDHC). For each hierarchical clustering, the tree was cut to create 1–5685 clusters. For a fixed number of clusters a pair of genes (from the 1044 genes assigned to at least one pathway) belonging to the same cluster was assumed to be a 'true positive' if the two genes both belonged to at least one specific KEGG pathway, and it was considered to be a 'false positive' if they did not share a single KEGG pathway. True and false positive rates were then obtained by dividing the number of true/false positives with the total number of gene pairs sharing a common KEGG pathway and total number of gene pairs not sharing a KEGG pathway respectively. When all genes are placed in their own individual clusters (5685 clusters), both true and false positive rates are equal to zero. As we reduce the number of clusters, both true and false positive rates increase defining a ROC curve. At the extreme when all genes are placed in the same cluster, both true and false

Downloaded from http://bioinformatics.oxfordjournals.org/ at Princeton University on October 9, 2014

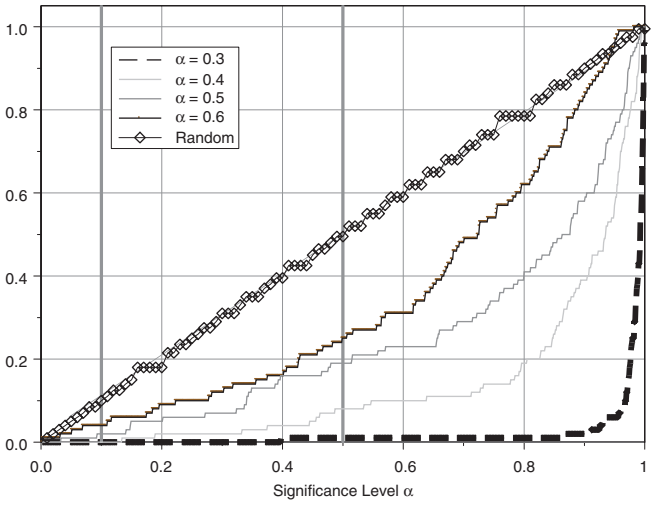

Fig. 3. Posterior probabilities as measures of statistical significance. The 'Random' scenario correspond to the situation in which all profiles were generated by the same multivariate normal probability distribution.

positive rates are equal to one. ROC curves derived in such a way for each dataset/method combination for the statistically relevant false positive rates (<0.05) are shown in Figure 4A. The global clustering methods, IMM and EDHC, both performed worse for the joint data analysis than using the cell-cycle data alone. The ROC curve for the CSIMM indicated that this method was able to integrate information from both datasets into a single more precise analysis. The behavior of the 'overfitted' model in which each microarray is treated as a separate context is inline with the behavior observed in the analysis of simulated data.

Owing to the strong imbalance between the total number of positive pairs 30 336 and negative pairs 513 067, a relatively low FPR still results in a large number of false positive pairs in comparison with the number of true positive pairs. Therefore, we examined more closely the behavior of different clustering procedures by relating the absolute number of false and true positive pairs of genes (Figure 4B). The improvements in precision of the CSIMM over competing approaches when looking at this outcome for <10 false positive pairs is dramatic.

Clusters of co-expressed genes at the highest ratio of true to false positives (191.6 with 5 false and 958 true positive pairs) along with corresponding KEGG pathways are displayed in Figure 5. The highest ratio of true to false positive was achieved when cutting the tree at the average linkage distance of 0.05. Genes were included in the heat map if they were assigned to at least one KEGG pathway and were co-clustered with at least one other gene from KEGG. The KEGG pathways implicated by these patterns are clearly related to the biological processes under investigation (sporulation and cell cycle). Although our 'gold standard' based on KEGG implicated five false positive pairs, a closer examination of the genes in two clusters on the top of the heat map (RNR1, MCD1, POL1, CDC45 and POL12) reveals that the activity of all these genes is tightly regulated during the DNA replication process. This indicates that at this level of resolution, CSIMM creates 'perfect' groupings of functionally related genes from KEGG. Interestingly, context-specific model for the cell-cycle data, in which gene expression profiles are split in two distinct cell-cycles performed better than the simple

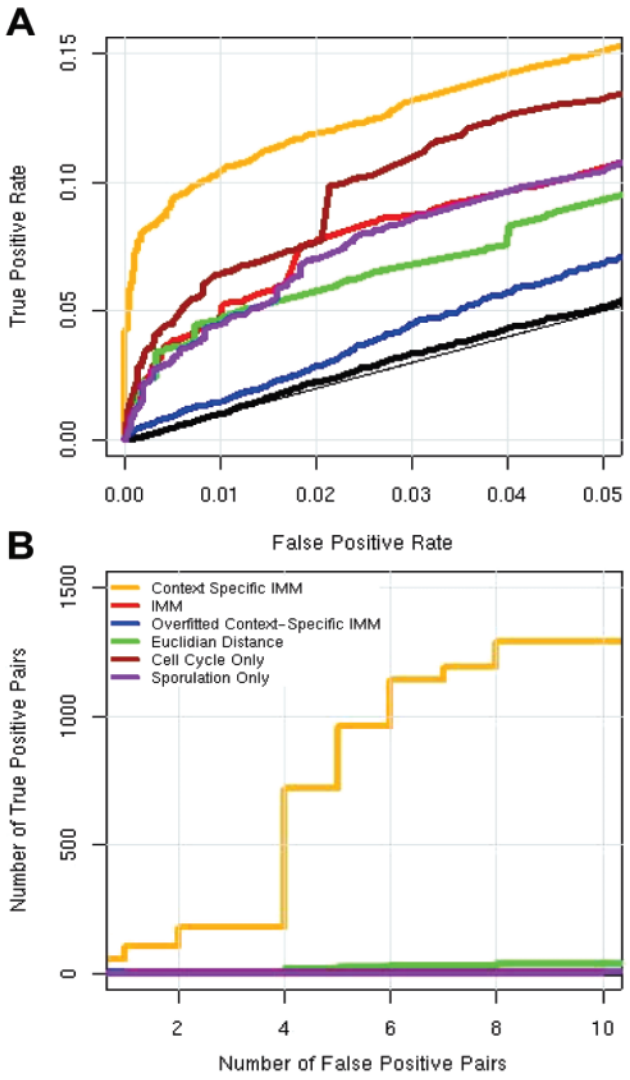

Fig. 4. ROC curves comparing the performance of different clustering approaches on the joint sporulation and cell cycle dataset. (**A**) The curve relating true positive and false positive rates. (**B**) The curve relating actual numbers of true positive and false positive pairs of co-clustered genes.

model when analyzing the cell-cycle data alone (Supplementary Figure S3). This could be a consequence of the issues previously raised about the synchronization of cells in different microarray experiments characterizing gene expression signatures of cell cycle (Cooper and Shedden, 2003).

We examined broader patterns of expression implicated at this level of significance by clustering all 135 genes that were co-clustered with at least one other gene after cutting the tree at the average linkage distance of 0.05 regardless of their KEGG membership (Supplementary Figure S4). In addition to KEGG pathway memberships, we examined correlations of the clusters generated by CSIMM with transcription factors shown to bind their promoters in 'Chip-on-Chip' experiments (Lee *et al*., 2002). The hierarchical tree in Supplementary Figure S4 was cut in eight clusters. Six of these clusters had more than two genes and they were tested for over-representation of genes whose promoters are substrates of any one

Downloaded from http://bioinformatics.oxfordjournals.org/ at Princeton University on October 9, 2014

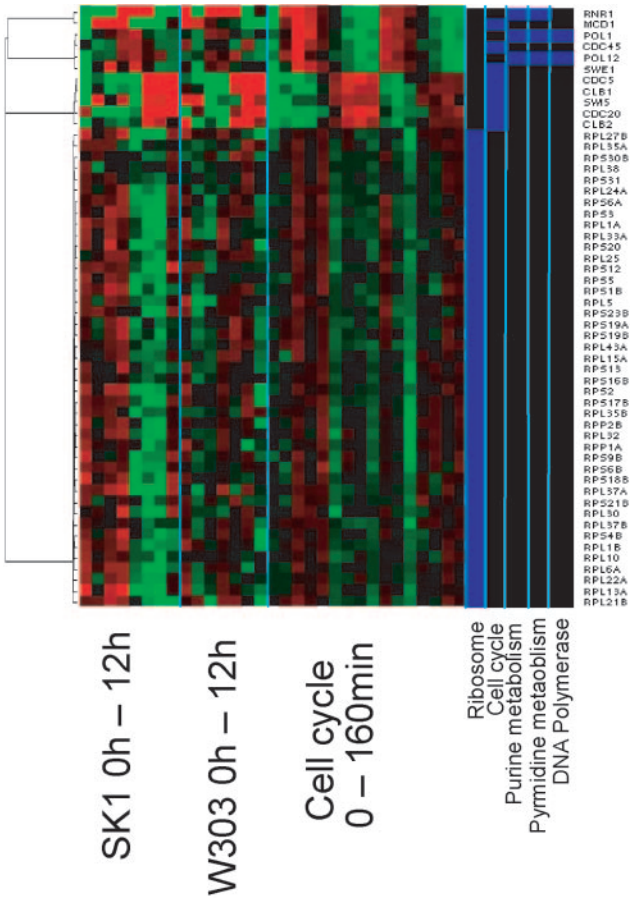

Fig. 5. Gene expression levels (green–red heat map) and KEGG pathways memberships (blue heat map) for 54 genes which were co-clustered with at least one other gene after cutting CSIMM-based tree at the average linkage distance of 0.05.

single transcription factor using the Fisher's exact test. Eight transcription factors were significantly associated with at least one of the cluster and their functional roles are closely related to the biological processes examined, as well as the KEGG pathway associations, landing additional credibility to the clusters identified by CSIMM. In comparison, cutting the tree formed by the Euclidian-distance based hierarchical clustering to obtain 135 genes that were co-clustered with at least one other gene generated diffused patterns without any obvious clustering structure (Supplementary Figure S5). Separate analyses of cell-cycle and sporulation data offered similar picture (Supplementary Figures S6–S9).

Finally, we investigated the validity of PPP-derived significance levels in deciding which clusters of genes are statistically significant. This was assessed by examining the proportion of false positive co-clusterings in clusters obtained by 'cutting' the hierarchical tree at different levels of average-linkage distances derived from posterior pairwise probabilities of co-expression. If the tree was cut at the similarity level d, the average PPP between each gene in a cluster and all other genes in the same cluster is greater than $(1 - d)$. At the same time the false discovery rates (FDR) are calculated as the proportion of implicated pairs of co-expressed genes when the tree is cut at the average-linkage distance d

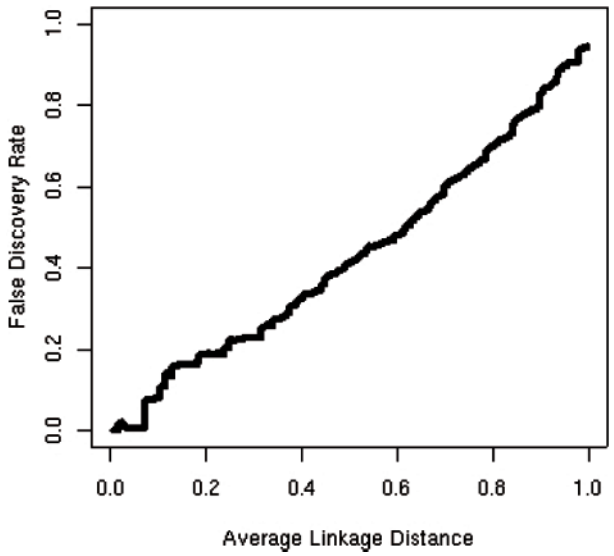

Fig. 6. Empirical FDR as a function of the average linkage distance used to cut the CSIMM-based hierarchical clustering tree.

which also shared at least one KEGG pathway out of all pairs implicated. Plotting the FDR against different ds (Figure 6) indicates that ds very well-approximate the empirical FDR.

4 DISCUSSION

The most important distinguishing feature of the model described here lies in its ability to circumvent the difficult problem of identifying the 'correct' number of local and/or global patterns in the data. Previously described context-specific models relied on different versions of penalized likelihood scores to estimate the 'correct' number of patterns in the data. There are some obvious advantages of being able to identify the single most likely number of clusters. However we previously demonstrated that our model-averaging results in more accurate analysis than the clustering procedure in which the 'correct' number of clusters is estimated from data. Here we further demonstrate that posterior distribution of clusterings offers a credible assessment of statistical significance of identified clusters and devise a practical approach for identifying statistically significant patterns in the data. This also simplifies the use of the model-based clustering since the whole procedure resembles simple hierarchical approaches.

The notion of context specificity introduced in our model is different from the two previously proposed context-specificity definitions. In the context-specific finite mixture model introduced by Barash and Friedman (2002), all 'uninformative' measurements within a context are placed into a single default cluster. CSIMM instead forms distinct groups of global clusters within each context. The module-network described by Segal *et al*. (2003) introduces a notion of context-specificity in which contexts are defined differently for different clusters and the distribution of all measurements within the same cluster and context are represented by the univariate Gaussian distribution. These two methods also facilitate estimation and modeling of the most likely context structure while CSIMM at this point requires context structure to be specified in advance. On

Downloaded from http://bioinformatics.oxfordjournals.org/ at Princeton University on October 9, 2014

the other hand, CSIMM uses globally defined contexts which are identical for all clusters, and the patterns within different contexts are described by multivariate Gaussian distributions. The distinction between univariate versus multivariate definition of local patterns seems to be particularly important in the situations when distinct local clusters describe complex patterns such as the time series or dose–response data.

ACKNOWLEDGEMENTS

The development of statistical models presented here has been supported by the grant 1R21HG002849 from NHGRI. Yeung is supported by NIH-NCI 1K25CA106988.

Conflict of Interest: none declared.

REFERENCES

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

Cho,R.J. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.

Cooper,S. and Shedden,K. (2003) Microarray analysis of gene expression during the cell cycle. *Cell Chromosome*, **2**, 1.

Cowell,R.G. *et al.* (1999) *Probabilistic Networks and Expert Systems.* Springer, New York.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Gelfand,E.A. and Smith,A.F.M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, **85**, 398–409.

Gelman,A. *et al.* (2003) *Bayesian Data Analysis*. CRC Press, New York.

Kanehisa,M. *et al.* (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.

Lee,T.I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

McLachlan,G.J. *et al.* (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413–422.

Medvedovic,M. and Guo,J. (2004) Bayesian model-averaging in unsupervised learning from microarray data. In *Proceedings of the fourth Workshop on Data Mining in Bioinformatics BIOKDD 2004,* Seattle, WA, USA, pp. 40–47.

Medvedovic,M. and Sivaganesan,S. (2002) Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, **18**, 1194–1206.

Medvedovic,M. *et al.* (2004) Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, **20**, 1222–1232.

Neal,R.M. (2000) Markov chain sampling methods for dirichlet process mixture models. *J. Comput. Graph. Stat.*, **9**, 249–265.

Primig,M. *et al.* (2000) The core meiotic transcriptome in budding yeasts. *Nat. Genet.*, **26**, 415–423.

Segal,E. *et al.* (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.

Segal,E. *et al.* (2004) A module map showing conditional activity of expression modules in cancer. *Nat. Genet.*, **36**, 1090–1098.

Stuart,J.M. *et al.* (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Yeung,K.Y. and Bumgarner,R. (2004) Pattern recognition in gene expression data. *Rec. Devel. Nucleic Acids Res.*, **1**, 333–354.

Downloaded from http://bioinformatics.oxfordjournals.org/ at Princeton University on October 9, 2014

APPENDIX: CSIMM MODEL

The statistical model describing the distribution of the data is given in the form of a Bayesian hierarchical model (Gelman *et al*., 2003). Dependencies between various model parameters and the data are defined by the Directed Acyclic Network (Cowell *et al*., 1999) in Figure A1. Nodes in the network represent random variables and arcs define the independence structure of the joint probability distribution function. Assuming that the probability distribution of any node is independent of its non-descendants if values of the parent nodes are given (Directed Markov Assumption), the joint probability distribution of all parameters and data is given by the product of the local probability distributions of individual random variables given their parents.

$$p(\mathbf{X},\mathbf{C},\mathbf{L},\mathbf{M},\mathbf{M}^*,\mathbf{S},\alpha,a,\lambda,\tau,\beta,\phi) = p(\mathbf{X}\,|\,\mathbf{C},\mathbf{M},\mathbf{S})p(\mathbf{C}\,|\,\alpha)p(\mathbf{M}\,|\,\mathbf{L},\mathbf{M}^*)p(\mathbf{S}\,|\,\beta,\phi)\; p(\mathbf{L}\,|\,\mathbf{C},a)p(\mathbf{M}^*\,|\,\lambda,\tau)p(\alpha)p(a)p(\lambda)p(\tau)p(\beta)p(\phi)$$

$\mathbf{M} = \{\mu_1,\ldots,\mu_Q\}$ is the set of all mean vectors associated with Q global patterns, $\mathbf{S} = \{\Sigma_1,\ldots,\Sigma_Q\}$ is the set of corresponding variance-covariance matrices, $\mathrm{M}^* = \{(\mu^*_{11},\ldots,\mu^*_{K_11}),\ldots,(\mu^*_{1R},\ldots,\mu^*_{K_RR})\}$ is the set of all local mean vectors, $\mathbf{S}^* = \{\Sigma^*_1,\ldots,\Sigma^*_R\}$ is the set of corresponding variance-covariance matrices and K_f is the number of local groupings of global clusters within context *f*. α, a, λ, τ, β and ϕ are hyperparameters for **C**, **L**, $\mathbf{M}^*$ and **S** respectively. The probability distribution of the expression data vector for gene i, given its classification variable c_i, global means **M** and the variance–covariance matrices **S** is

$$p(x_i\,|\,c_i = q,\mathbf{M},\mathbf{S}) = f_N(x_i\,|\,\mu_q,\Sigma_q),$$

where $f_N(.\,|\,\mu,\boldsymbol{\Sigma})$ is the multivariate Gaussian probability distribution function with mean μ and variance–covariance matrix $\boldsymbol{\Sigma}$. All variance–covariance matrices in the model are context-specific and diagonal. That is Σ_q is the block diagonal matrix with context-specific diagonal matrices $\sigma^2_{tf}\mathbf{I}$ on the diagonal.

The probability distribution of the global mean vector μ_q, given the local structure **L** and the local parameters $\mathbf{M}^*$ and $\mathbf{S}^*$ is

$$p((\mu_{q1},\mu_{q2},\ldots,\mu_{q\mathrm{R}})\,|\,\mathbf{L},\mathbf{M}^*,\sigma^*) = f_N(\mu_{q1}\,|\,\mu^*_{L_{q1}1},\Sigma^*_1)f_N(\mu_{q2}\,|\,\mu^*_{L_{q2}2},\Sigma^*_2)\cdots f_N(\mu_{qR}\,|\,\mu^*_{L_{qR}R},\Sigma^*_R),$$

where μ_{qf} is the subvector of the global mean μ_q on the *f*-th context. Prior distributions for the local groupings **L** are defined following the infinite mixtures approach that avoids the specification of the 'correct' number of groups of local clusters for each context (Medvedovic *et al*., 2004; Medvedovic and Sivaganesan, 2002). The probability of assigning the global cluster q to an already existing group of clusters t within the context f, given **C** and a, is given by $p(L_{qf} = t\,|\,C,a) \propto \frac{n_{-qft}}{Q-1+a}$, $t = 1,\ldots,\mathrm{Q}$, where n_{-qft} is the number of global clusters currently placed in local cluster t within context f without counting global cluster q. The probability of assigning a global cluster to a new local group is given by

$$p(L_{qf} \neq L_{qtf},\forall q' \neq q\,|\,C,a) \propto \frac{a}{Q-1+a}.$$

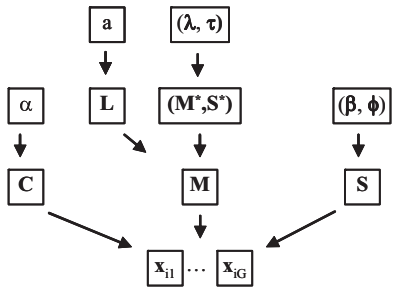

Fig. A1. Context-specific infinite mixtures.

The rest of the local conditional probability distributions, the structure of variance–covariance matrices and hyperparameters are stated in the Supplementary Material.

The joint posterior distribution of all parameters in the model given data is estimated using Gibbs sampler. Gibbs sampler (Gelfand and Smith, 1990) is a general procedure for sampling observations from a multivariate distribution. It proceeds by iteratively drawing observations from complete conditional distributions of all components given the current values of all other components. Under mild condition, the distribution of generated multivariate observations converges to the target multivariate distribution. The Gibbs sampler employed here is derived from previously described algorithms for fitting infinite mixture models. Conditional posterior distributions for **M**, $\mathbf{M}^*$ and **L** are derived assuming that $\Sigma^*_f = (\sigma^*_q)^2\mathbf{I}$ and by letting $\sigma^*_q \to 0$ within all contexts. This effectively forces all global cluster means grouped together within a context to be identical within this context. Consequently, instead of estimating means and variances for each of the Q global clusters within each context f, we estimate only $K_f < Q$ local parameters.

The posterior distributions for the local classification variables, conditional on all other parameters in the model are

$$p(L_{qf} = t\,|\,C,X,a) \propto \frac{n_{-qft}}{Q-1+a} f_N(\bar{\mathbf{x}}^q_f\,|\,\mu^*_{tf},\frac{\sigma^2_{tf}}{n_q}\mathbf{I})$$

$$p(L_{qf} \neq L_{qtf},\forall q' \neq q\,|\,C,a) \propto \frac{a}{Q-1+a}\int f_N(\bar{\mathbf{x}}^q_f\,|\,\boldsymbol{\mu}^*_{tf},\frac{\sigma^2_{tf}}{n_q}\mathbf{I})\,p(\boldsymbol{\mu}^*_{ft},\frac{\sigma^2_{tf}}{n_q})\mathrm{d}(\boldsymbol{\mu}^*_{tf},\sigma^2_{tf}),$$

where $\bar{\mathbf{x}}^q_f = \frac{\sum_{c_i=q}\mathbf{x}^f_i}{n_q}$

All other conditional posterior distributions are similar to the simple infinite mixture models (Medvedovic *et al*,. 2004; Medvedovic and Sivaganesan, 2002) and are given in the web Supplementary Material. The Gibbs sampler is initialized sampling all model parameters from their respective prior distributions, and placing all global gene expression profiles into a single cluster. The Gibbs sampler proceeds to sample first global clusters, then local groupings of global clusters within each context and then the rest of the parameters in the model. To alleviate the problem of 'slow mixing', we apply heuristic annealing adjustment described in the web Supplementary Material. Previously, we demonstrated that such modifications preserve the topology of the posterior distribution of clusterings (Medvedovic *et al*., 2004).

Downloaded from http://bioinformatics.oxfordjournals.org/ at Princeton University on October 9, 2014