



University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

This paper is made available online in accordance with publisher policies. Please scroll down to view the document itself. Please refer to the repository record for this item and our policy information available from the repository home page for further information.

To see the final version of this paper please visit the publisher's website. Access to the published version may require a subscription.

Author(s): Carl Edward Rasmussen, Bernard J. de la Cruz, Zoubin Ghahramani, and David L. Wild

Article Title: Modeling and Visualizing Uncertainty in Gene Expression Clusters Using Dirichlet Process Mixtures

Year of publication: 2009

Link to published version:

<http://dx.doi.org/10.1109/TCBB.2007.70269>

Publisher statement: "©2009 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE

Modeling and Visualizing Uncertainty in Gene Expression Clusters Using Dirichlet Process Mixtures

Carl Edward Rasmussen, Bernard J. de la Cruz, Zoubin Ghahramani, and David L. Wild

Abstract—Although the use of clustering methods has rapidly become one of the standard computational approaches in the literature of microarray gene expression data, little attention has been paid to uncertainty in the results obtained. Dirichlet process mixture (DPM) models provide a nonparametric Bayesian alternative to the bootstrap approach to modeling uncertainty in gene expression clustering. Most previously published applications of Bayesian model-based clustering methods have been to short time series data. In this paper, we present a case study of the application of nonparametric Bayesian clustering methods to the clustering of high-dimensional nontime series gene expression data using full Gaussian covariances. We use the probability that two genes belong to the same cluster in a DPM model as a measure of the similarity of these gene expression profiles. Conversely, this probability can be used to define a dissimilarity measure, which, for the purposes of visualization, can be input to one of the standard linkage algorithms used for hierarchical clustering. Biologically plausible results are obtained from the Rosetta compendium of expression profiles which extend previously published cluster analyses of this data.

Index Terms—Clustering, classification, and association rules, biology and genetics, bioinformatics (genome or protein) databases, statistical computing, stochastic processes, Monte Carlo.



1 INTRODUCTION

THE use of clustering methods has rapidly become one of the standard computational approaches to understanding microarray gene expression data [1], [2], [3]. In clustering, the patterns of expression of different genes across time, treatments, and tissues are grouped into distinct clusters (perhaps organized hierarchically) in which genes in the same cluster are assumed to be potentially functionally related or to be influenced by a common upstream factor. Such cluster structure can be used to aid the elucidation of regulatory networks. For example, a compendium of gene expression profiles corresponding to mutants and chemical treatments can be used as a systematic tool to identify gene functions because mutants or drug targets that display similar profiles are likely to share cellular functions [4]. It would also be expected that gene knockouts/mutations or treatments that have impact on the same signaling or metabolic pathway or affect the same organelle would exhibit some overlap in altered gene expression profiles.

Agglomerative hierarchical clustering [1] is one of the most frequently used methods for clustering gene expression

profiles. However, commonly used methods for agglomerative hierarchical clustering rely on the setting of some score threshold to distinguish members of a particular cluster from nonmembers, making the determination of the number of clusters arbitrary and subjective. The algorithm provides no guide to choosing the “correct” number of clusters or the level at which to prune the tree. It is often difficult to know which distance metric to choose, especially for structured data such as gene expression profiles. Moreover, these approaches do not provide a measure of uncertainty about the clustering, making it difficult to compute the predictive quality of the clustering and to make comparisons between clusterings based on different model assumptions (e.g., numbers of clusters and shapes of clusters). In this paper, we use statistical inference to overcome these limitations. An important issue that must be addressed in any clustering method is the question of how many clusters to use. Bayesian statistics and model-based approaches can provide elegant solutions to model selection questions of this kind. With these approaches, there is no need to make arbitrary choices about how many clusters there are in the data; nevertheless, after modeling, one can still ask questions such as “how probable is it that two genes belong to the same cluster?”

Within a Bayesian framework, all assumptions are presented in terms of priors and the choice of likelihood function. Since it seems unreasonable to assume that complex gene expression data have been generated by some small finite number of causes, an elegant nonparametric approach is to assume that the data was in fact generated from an *infinite number of Gaussian clusters*. In a Gaussian clustering model, each gene expression profile represents a multidimensional vector of measurements and the probability distribution for each cluster is assumed to be

- C.E. Rasmussen and Z. Ghahramani are with the Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, UK. E-mail: cer54@cam.ac.uk, zoubin@eng.cam.ac.uk.
- B.J. de la Cruz is with the Beckman Research Institute, City of Hope National Medical Center, 1500 E. Duarte Road, Duarte, CA 91010. E-mail: bjdelacruz@gmail.com.
- D.L. Wild is with the Warwick Systems Biology Centre, University of Warwick, Coventry House, Coventry, CV4 7AL, UK. E-mail: d.l.wild@warwick.ac.uk.

Manuscript received 8 June 2007; revised 8 Oct. 2007; accepted 25 Oct. 2007; published online 30 Nov. 2007.

For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2007-06-0067. Digital Object Identifier no. 10.1109/TCBB.2007.70269.

a multivariate Gaussian. We describe an approach to the problem of automatically clustering microarray gene expression profiles based on the theory of infinite Gaussian mixtures (or Dirichlet process mixtures, DPM) [5], [6]. This theory is based on the observation that the mathematical limit of an infinite number of components in an ordinary finite mixture model (i.e., clustering model) corresponds to a Dirichlet process prior [5], [6], [7]. In an infinite Gaussian mixture model, there is no need to make arbitrary choices about how many clusters there are in the data. Although in theory, the infinite mixture model has an infinite number of parameters, surprisingly, it is possible to do exact inference in these infinite mixture models efficiently using Markov chain Monte Carlo (MCMC) methodology, since only the parameters of a finite number of the mixture components need to be represented explicitly. The theory of DPM models has a long history, going back to [7], [8], and [9], and has recently become popular with the availability of fast MCMC inference, see [6] and [10] for early examples. We first proposed and implemented the application of DPMs to clustering gene expression profiles in an extended conference abstract in 2002 [11]. Although this work is not widely known and cited, many groups have subsequently independently rediscovered the value of a fully Bayesian analysis based on DPMs to this problem [12], [13], [14], [15], [16]. We have also subsequently applied the approach to the clustering of protein sequences [17].

In this paper, we illustrate our methods in detail, with a practical application to a well-studied data set: the Rosetta compendium of expression profiles corresponding to 300 diverse mutations and chemical treatments in *S. cerevisiae* [4]. We describe a simple but novel method of visualizing the results which facilitates comparison with the dendrograms obtained by the usual hierarchical clustering approach to this type of data. While our results confirm many of the previously published clusters identified in this data set, they also provide new biological insights by revealing a finer level of granularity in the clustering. These results are consistent with recent literature which suggests that distinct functions may share proteins and have overlapping regulatory mechanisms.

2 METHODS

2.1 Dirichlet Process Mixture Models

Although hierarchical clustering is the most widely used method for clustering gene expression data, model-based nonhierarchical methods have also been explored. One commonly used computational method of nonhierarchical clustering based on measuring euclidean distance between gene expression profiles is given by the k-means algorithm [18], [19]. However, the k-means algorithm is inadequate for describing clusters of unequal size or shape [20]. A generalization of k-means can be derived from the theory of maximum likelihood estimation of Gaussian mixture models [21], [22]. In a Gaussian mixture model, the data (e.g., gene expression profiles, which can be arranged into p -dimensional vectors \mathbf{y}) is assumed to have been generated from a finite number (k) of Gaussians:

$$P(\mathbf{y}) = \sum_{j=1}^k \phi_j P_j(\mathbf{y}), \quad (1)$$

where ϕ_j is the mixing proportion for cluster j (fraction of population belonging to cluster j ; $\sum_j \phi_j = 1$; $\phi_j \geq 0$), and $P_j(\mathbf{y})$ is a multivariate Gaussian distribution with mean μ_j and covariance matrix Σ_j . The clusters can be found by fitting the maximum likelihood Gaussian mixture model as a function of the set of parameters $\theta = \{\phi_j, \mu_j, \Sigma_j\}_{j=1}^k$ using the EM algorithm [21]. Euclidean distance corresponds to assuming that the Σ_j are all equal multiples of the identity matrix.

Starting from a finite mixture model (1), we define a prior over the mixing proportion parameters ϕ . The natural conjugate prior for mixing proportions is the symmetric Dirichlet distribution, with concentration parameter α/k :

$$P(\phi|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \prod_{j=1}^k \phi_j^{\alpha/k-1}, \quad (2)$$

where α controls the distribution of the prior weight assigned to each cluster, and Γ is the gamma function.

We then explicitly include indicator variables c_i for each data point (i.e., gene expression profile) which can take on integer values $c_i = j$, $j \in \{1, \dots, k\}$, corresponding to the hypothesis that data point i belongs to cluster j . Under the mixture model, by definition, the prior probability is proportional to the mixing proportion: $P(c_i = j|\phi) = \phi_j$. A key observation is that we can compute the conditional probability of one indicator variable given the setting of all the other indicator variables after *integrating over* all possible settings of the mixing proportion parameters:

$$\begin{aligned} P(c_i = j|\mathbf{c}_{-i}, \alpha) &= \int P(c_i = j|\mathbf{c}_{-i}, \phi) P(\phi|\mathbf{c}_{-i}, \alpha) d\phi \\ &= \frac{n_{-i,j} + \alpha/k}{n - 1 + \alpha}, \end{aligned} \quad (3)$$

where \mathbf{c}_{-i} is the setting of all indicator variables except the i th, n is the total number of data points, and $n_{-i,j}$ is the number of data points belonging to cluster j not including i . By Bayes rule

$$P(\phi|\mathbf{c}_{-i}, \alpha) = P(\phi|\alpha) / P(\mathbf{c}_{-i}|\alpha) \prod_{l \neq i} P(c_l|\phi), \quad (4)$$

which is also a Dirichlet distribution, making it possible to perform the above integral analytically. We can now take the limit of k going to infinity, obtaining a Dirichlet process with differing conditional probabilities for clusters with and without data: For clusters where $n_{-i,j} > 0$: $p(c_i = j|\mathbf{c}_{-i}, \alpha) = \frac{n_{-i,j}}{n-1+\alpha}$. For all other clusters combined: $p(c_i \neq c_{i'} \text{ for all } i' \neq i|\mathbf{c}_{-i}, \alpha) = \frac{\alpha}{n-1+\alpha}$. This shows that the probabilities are proportional to the occupation numbers, $n_{-i,j}$. Using these conditional probabilities, one can Gibbs sample from the indicator variables efficiently, even though the model has infinitely many Gaussian clusters. Having integrated out the mixing proportions, one can also Gibbs sample from all of the remaining parameters of the model, i.e., $\{\mu, \Sigma\}_j$, or one can integrate these out as well. The details of these procedures can be found in [6].

2.2 Data Preprocessing

All gene expression profile data were obtained from the web site http://www.rii.com/tech/pubs/cell_hughes.htm. Data from the treatment and mutant experiments were concatenated with the control ("wild-type") experiments. To facilitate direct comparison of our results with previously published work, profiles were selected from the raw data to include only experiments with two or more genes up or down-regulated by more than threefold, and significant at $p \leq 0.01$ under a gene-specific error model, as described by Hughes et al. [4]; and to include only genes that were up or down-regulated more than threefold, significant at $p \leq 0.01$ in two or more experiments. Following Hughes et al. [4], missing data were replaced by row (column) means.¹ The final data set comprised 636 genes and 194 experiments (including controls).

2.3 Computational Experiments

For all data sets, the dimensionality of the data was first reduced by projecting the data onto the 10 leading eigen-directions of the correlation coefficient matrix. These 10 directions captured most of the variance in the data. This 10-dimensional projection of the data, y , was then modeled with the DPM model. A fully Bayesian approach to choosing the number of dimensions of the low-dimensional projection is beyond the scope of this paper; however, one possibility would be based on defining a DPM of factor analyzers, which combines clustering with dimensionality reduction [23]. We have experimented with using 5 and 15 directions in the projection; in both cases, the inference algorithm discovers fewer represented mixture components.

The parameters of the model were assigned prior distributions following [6]. The priors on the parameters of the Gaussian mixtures were *conditionally conjugate*, specifically Gaussian for the means and Wishart for the covariances (with top-level parameters set to the moments of the data, such that the entire procedure is insensitive to translation, rotation, and rescaling of the data). The prior on the concentration parameter was chosen to be vague, identical with [6].

The mixture model was initialized with all data belonging to a single Gaussian, and a large number of Gibbs sampling sweeps are performed, updating all variables and parameters, i.e., $\{\{\mu_j, \Sigma_j\}, \{c_i\}, \alpha\}$, in turn by sampling from the conditional distributions derived in the previous sections and described in more detail in [6]. To assess the mixing time, we examined the autocorrelation coefficients for the number of represented components (see Fig. 1). We chose the number of represented components as a diagnostic, as this is one of the properties of the state which changes most slowly. We estimated the mixing time as the sum of the autocorrelation coefficients from a large negative lag to large positive lag. For the transcript response clustering experiment, shown in Fig. 1, the mixing time is about 200. We then ran the final MCMC to generate 100 roughly independent samples, by using a burn-in of 10,000 samples, and then saving every

1. We note that a full Bayesian treatment of missing data would involve integrating over the missing values.

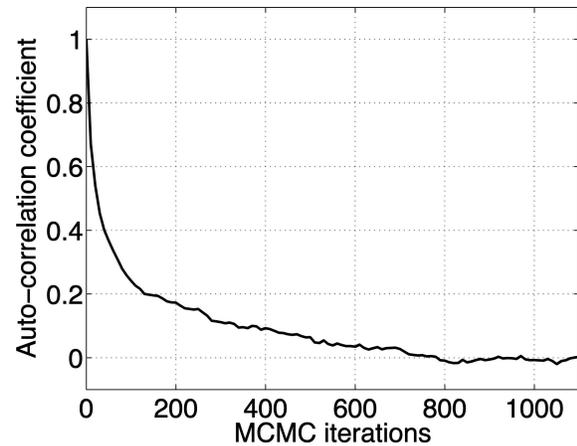


Fig. 1. The autocorrelation coefficient function for the number of represented components for the clustering of gene experiments. The function is only shown for positive lags, but is symmetric. The area under the curve (including both sides) is about 200.

1,000th sample for the next 100,000 samples. This took 34 minutes on a desktop computer. For the clustering of experimental conditions, a similar strategy reveals a somewhat slower mixing time of 60,000. We thus ran the chain initially for 100,000 iterations for burn-in, and then for 11,000,000 samples, keeping every 100,000th to get 100 roughly independent samples. This takes about 11 hours on a desktop, but the results of a 100 times shorter run (6 minutes) are virtually indistinguishable.

2.4 Visualization of Results

We wish to determine the probability that two genes belong to the same cluster, i.e., have similar functional roles or are influenced by a common upstream factor. Unlike methods based on a single clustering of the data, the approach described in this paper computes this probability while taking into account all sources of model uncertainty (including number of clusters and location of clusters). Specifically, we use the probability p_{ij} that two genes i and j belong to the same cluster in the DPM model as a measure of the similarity of these gene expression profiles. Conversely, $1 - p_{ij}$ defines a *dissimilarity measure*, which for the purposes of visualization, can be input to one of the standard linkage algorithms used for hierarchical clustering (Fig. 6). We can easily compare the dendrograms thus obtained to the usual hierarchical clustering approach, which computes distance metrics directly on the gene expression profiles or correlation coefficients between profiles [1]. Clustering is done in both directions: both by gene transcripts and by experimental profiles.

2.5 Annotation of Clusters by Gene Ontology

An important first step toward obtaining a functional profile of a gene list is to cluster the genes in terms of a comprehensive, well-structured set of functional categories such as that defined by the Gene Ontology (GO) Database. GO provides three structured ontologies of defined terms to describe gene product attributes: biological process, molecular function, and cell component. Groups annotated at the highest level in the GO hierarchy (biological process) are likely to contain genes involved in related pathways. In

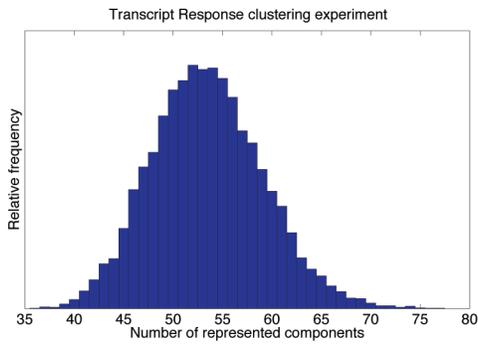


Fig. 2. Number of represented components for the clustering of transcript responses experiment. This figure shows that the relative number of components along the MCMC run varies between about 40 and 70.

order to find statistically significant GO annotations related to a given cluster of genes, we looked for annotation terms that are overrepresented in this cluster. The probability that this overrepresentation is not found by chance can be calculated by the use of a hypergeometric test. Because of the effects of multiple testing, a subsequent correction of the p -values is necessary, and we used the SGD GO Term Finder <http://db.yeastgenome.org/cgi-bin/GO/goTermFinder> [24], which applies a Bonferroni correction.

After identifying clusters and their members, the SGD GO Term Finder was used to determine whether clusters were overrepresented by particular cellular localization, molecular function, or molecular process GO terms. Absolute p -value depends on size of clusters and the size of the reference list, in this case, all yeast ORFs with an assigned GO term. The set of experimental clusters shrinks when we exclude double mutants, chemical treatments, and wild type profiles. It should also be noted that SGD GO Term Finder does not calculate underrepresented GO terms and this has not been considered here. It can be seen for some clusters that the assigned GO term may be either too specific or too general. For example, cluster 15 of the clustering of experiments has as its top molecular process GO term “physiological process,” a high-level ontology but not insightful. For the same cluster, the best molecular function GO term is given as “hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides”—this is a low-level, highly specific function yet also not immediately insightful. Rather than focusing on the best hit alone, all significant GO terms are used to provide insight (see supplemental material, Tables 1, 2, 3, 4, 5, and 6, which can be found on the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TCBB.2007.70269>).

3 RESULTS AND DISCUSSION

3.1 Clustering by Transcript Response

In all, 636 transcripts were found to meet the prefiltering criteria described in Section 2. That is, these genes are those most affected by the gene knockouts/treatments which constitute the experimental conditions. In Fig. 2, we show the relative frequency of the number of represented components over the MCMC samples. It shows that

between 40 and 70 components are likely. This wide range of number of clusters underlines our premise that the individual clusterings found are associated with substantial uncertainties. Rather than picking one particular clustering, in the following, we always visualize properties averaged over all states sampled by MCMC.

In Fig. 3, we show the number of times, out of 100 samples, that the *indicator variables* for two genes were equal. As described in Section 2, this may be interpreted as the probability p_{ij} that two genes i and j belong to the same cluster, and the different colors represent this probability. We refer to p_{ij} as the *co-occurrence probability* of genes i and j . The granularity of this clustering is determined by the data and not by some user-defined threshold. Large solid blocks of color along the diagonal correspond to homogeneous clusters. Note that in our method, sequences may partially belong to more than one cluster; off-diagonal elements indicate such “cross-clustering” or overlapping clusters. These off-diagonal blocks (such as cluster 2 or 4 in Fig. 3) may indicate one of two possibilities: It may mean that there is *uncertainty* in whether a set of genes should be assigned to one of the two clusters, or it may indicate a set of genes which should really belong *simultaneously* to two clusters. In this latter case, the fundamental assumption that a gene belongs to only one cluster does not apply, and suggests the existence of overlapping regulatory pathways. We focus on 17 transcript response clusters (TCs) represented as blocks of color along the diagonal (cluster members are given in Tables 1 and 2). Of these, 11 clusters form a single group along the diagonal; while in five cases, the clusters are broken into subclusters (clusters 2, 4, 9, 12, and 15). These are seen as mirrored bands above the background color (dark blue) and off the diagonal. The subclusters indicate that, while their members are most closely linked, there is also simultaneously a weaker affinity for other clusters. Using the SGD GO Term Finder, we determined overrepresented GO terms for each of the 17 TCs. The top GO term and the p -value for each TC is given in Tables 3, 4, and 5. Significance is defined as $p < 10^{-2}$.

Hughes et al. [4] applied agglomerative hierarchical clustering using a correlation coefficient-based distance metric [1]. They identified eight main transcript response clusters: PAU, RNR2,3,4, ergosterol, amino acid biosynthesis, calcineurin/PKC, mitochondrial function, mating, and S/C (general stress response and carbohydrate metabolism). The PAU cluster includes a family of genes noted only for their lack of serine residues, and for being induced during anaerobic growth, but which otherwise do not have a known function [25], [26]. The RNR cluster represents genes that respond to DNA damage. The following TCs in Fig. 3 appear to match with the following groups found by Hughes et al.: PAU (TC 5), RNR (TC 3), ergosterol (TC 13), mitochondrial function (TC 1), and mating (TC 6). The other clusters described by Hughes et al., in particular, the S/C cluster and amino acid biosynthesis cluster, are distributed over several TC clusters. In particular, TC 4 (monosaccharide transport), TC 7 (general amino acid biosynthesis), TC 8 (carbohydrate transport), TC 14 (sulfur metabolism), TC 15 (vitamin metabolism), TC 16 (beta-alanine biosynthesis),

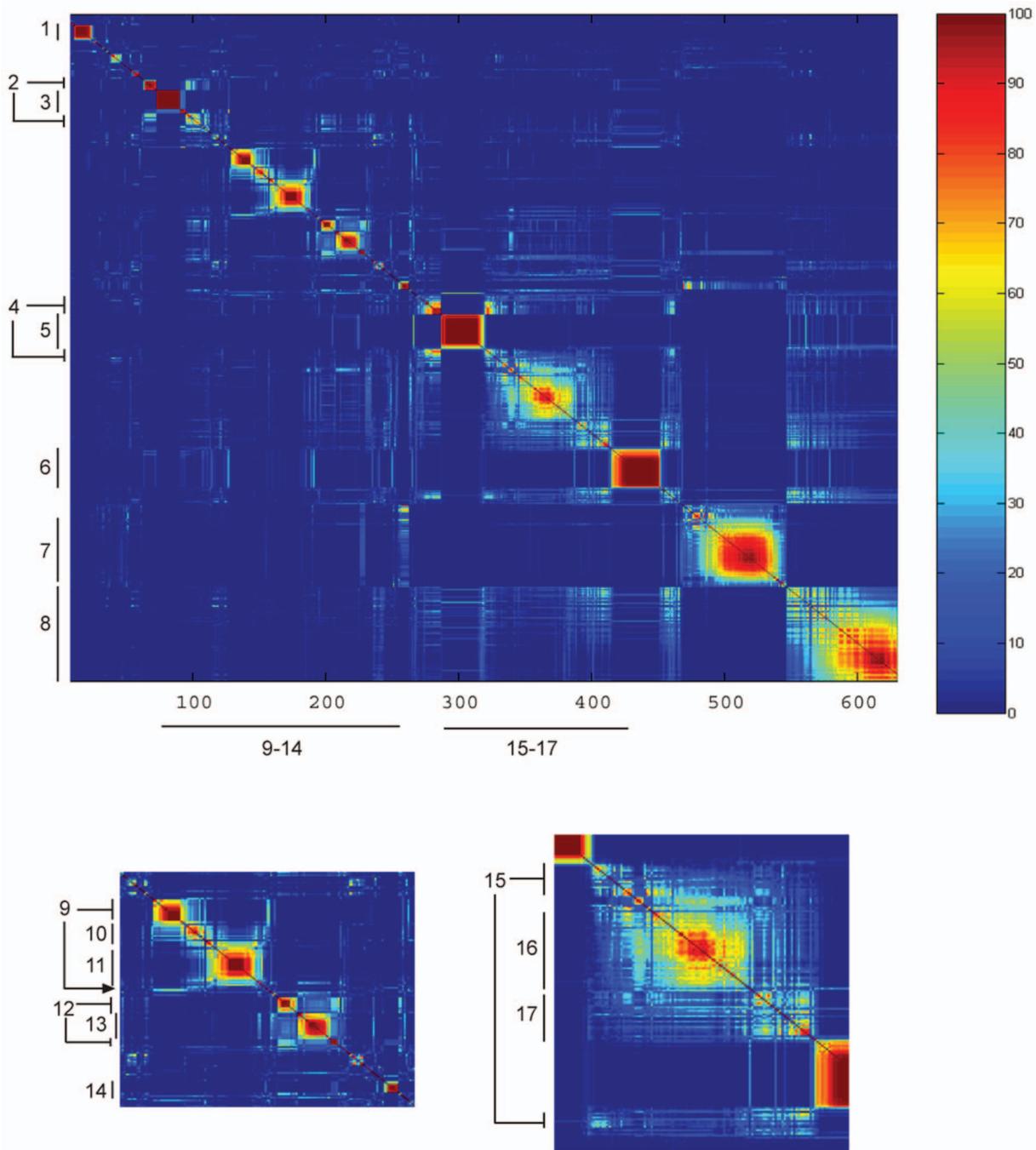


Fig. 3. Co-occurrence probabilities of the 636 transcript response clusters. This figure shows the number of times, out of 100 samples, that the *indicator variables* for two genes were equal. This may be interpreted as the probability p_{ij} that two genes i and j belong to the same cluster, and the different colors represent this probability. Numbers 1-17 indicated in the margins refer to the TCs discussed in detail in the text. The subfigures represent a magnified view of portions of the larger figure. A larger version of this figure is available in the supplementary materials, which can be found on the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TCBB.2007.70269>.

and TC 17 (polyamine transport). As such, the DPM method was able to distribute the general S/C and amino acid biosynthesis groups into more specific clusters.

TC 2, TC 11, and TC 12 all exhibit significance for "cell wall," "plasma membrane," and "cytokinesis" GO terms. Examination of the cluster members suggests that TC 2 is involved in the formation of the mating bud. The best process GO term associated with TC 11 is "cytokinesis, completion of separation." TC 12 is associated with process GO term "cell

wall organization and biogenesis." We note that for TC 5, the best hit for all three GO categories is "unknown." Cluster 5 is a large group (32 transcripts) and contains 20 out of 21 PAU genes (PAU 7 appears in TC 8). TC 5 also contains five DAN/TIR mannoprotein genes, which are typically part of the cell wall. This is in agreement with work indicating the importance of these sets of ORFs in cell wall integrity [27], suggesting that TC 5 is yet another "cell wall" cluster. This identification of a new cluster of "cell wall" transcripts

TABLE 1
List of ORFs in TCs

Cluster 1		Cluster 3 (continued)		Cluster 5 (continued)		Cluster 7	
YPR002W	PDH1	YOL106W		YLR037C	DAN2	YOR107W	RGS2
YGR035C		YER160C		YFL020C	PAU5	YLR237W	THI7
YMR102C		YLR334C		YBR301W	DAN3	YNR064C	
YOR136W	IDH2	YML045W		YGL261C	PAU11	YFR047C	BNA6
YDR406W	PDR15	YJR029W		YOL161C	PAU20	YBR045C	GIP1
YOR135C		YJR027W		YDR542W	PAU10	YKL218C	SRY1
YLR304C	ACO1	YHR214C-B		YHL046C	PAU13	YJL217W	
YOR153W	PDR5	YML039W		YNR076W	PAU6	YIL164C	NIT1
YNL037C	IDH1	YLR035C-A		YMR325W	PAU19	YBR294W	SUL1
YLR346C		YFL002W-A		YOR009W	TIR4	YOL064C	MET22
YOR049C	RSB1	YDR170W-A		YAL068C	PAU8	YJL089W	SIP4
YAL061W		YHR213W		YGR213C	RTA1	YGL009C	LEU1
YBL043W	ECM13			YJL114W		YNR069C	BSC5
YNR056C	BIO5			YIL175W		YER081W	SER3
		Cluster 4		YMR316C-B		YPL135W	ISU1
Cluster 2		YDL037C	BSC1	YGR144W	THI4	YBR105C	VID24
YLR042C		YDL039C	PRM7			YIL056W	VHR1
YOR247W	SRL1	YDL038C		Cluster 6		YNR058W	BIO3
YKR013W	PRY2	YCR021C	HSP30	YGL183C	MND1	YOR130C	ORT1
YPL163C	SVS1	YDR516C	EMI2	YNL180C	RHO5	YJR130C	STR2
YIL123W	SIM1	YER066C-A		YGR040W	KSS1	YGL180W	ATG1
YPL256C	CLN2	YDR343C	HXT6	YKL178C	STE3	YKL120W	OAC1
YJL158C	CIS3	YFR053C	HXK1	YHR145C		YLR162W	
YOR248W		YDR342C	HXT7	YOL104C	NDJ1	YKL121W	
YGR014W	MSB2	YER067W		YLR040C		YOR303W	CPA1
YDR309C	GIC2	YPR160W	GPH1	YJR004C	SAG1	YDL170W	UGA3
YGR189C	CRH1	YOL150C		YPL192C	PRM3	YJR154W	
YLR194C		YFL060C	SNO3	YIL082W		YOR337W	TEA1
YKR091W	SRL3	YEL011W	GLC3	YCL027W	FUS1	YOR339C	UBC11
YKR061W	KTR2	YBR183W	YPC1	YCL055W	KAR4	YHR208W	BAT1
YDR077W	SED1	YBL049W	MOH1	YDR124W		YGR239C	PEX21
YPL067C		YPL230W		YHR005C	GPA1	YIL165C	
YHR030C	SLT2	YDR277C	MTH1	YMR065W	KAR5	YGL125W	MET13
YLR121C	YPS3	YBL064C	PRX1	YIL015W	BAR1	YJR155W	AAD10
YPR078C		YML128C	MSC1	YLR452C	SST2	YER091C	MET6
YHR209W		YLL026W		YNR044W	AGA1	YDL198C	GGC1
YDR085C	AFR1	HSP104		YBL016W	FUS3	YNL104C	LEU4
YNL034W				YJL157C	FAR1	YDR127W	ARO1
YJL027C		Cluster 5		YGL032C	AGA2	YJR109C	CPA2
YGR156W	PTI1	YHR092C	HXT4	YCR089W	FIG2	YOL140W	ARG8
YKL163W	PIR3	YBR066C	NRG2	YBR083W	TEC1	YNR050C	LYS9
YOL011W	PLB3	YLL025W	PAU17	YCLX07W		YHR162W	
YEL021W	URA3	YDR213W	UPC2	YDR461W	MFA1	YER024W	YAT2
YLR391W-A		YOR394W	PAU21	YFL026W	STE2	YLR267W	BOP2
		YJR150C	DAN1	YKL209C	STE6	YER073W	ALD5
Cluster 3		YPL282C	PAU22	YCRX18C		YHR029C	YHI9
YLR343W	GAS2	YHR139C	SPS100	YML048W-A		YBR248C	HIS7
YBL101W-B		YJL223C	PAU1	YNL279W	PRM1	YDR158W	HOM2
YCL019W		YIR041W	PAU15	YNL145W	MFA2	YOR203W	
YIL060W		YKL224C	PAU16	YMR232W	FUS2	YLR152C	
YBL005W-B		YCR104W	PAU3	YIL080W		YDR035W	ARO3
YAR009C		YLL064C	PAU18	YIL082W-A		YMR097C	MTG1
YER138C		YEL049W	PAU2	YIL037C	PRM2	YJR111C	
YBR012W-B		YGR294W	PAU12	YJL170C	ASG7	YMR108W	ILV2
YMR050C		YIL176C	PAU14	YIL011W	TIR3	YPL250C	ICY2
YMR045C		YOR010C	TIR2	YBR250W		YER052C	HOM3

makes sense in light of the clustering of experimental conditions described below. While Hughes et al. identified a group of profiles collectively related to “cell wall,” the DPM clustering suggests that this large group forms smaller, distinctly regulated subclusters. Recent literature looking at cell wall proteins suggests that distinct functions—for example, controlling osmotic pressure, responding to physical stress, maintaining cell wall integrity, and providing a protein scaffold—may share proteins and have overlapping regulatory mechanisms [27]. Furthermore, the signaling pathways involve cross talk among MAPK kinase pathways [28]. For example, sets of cell wall proteins, such as the PAU family, are activated by pheromone signaling, by global

stress signaling, as well as the calcineurin-mediated signaling, suggesting multiple modes of regulation.

Likewise, rather than finding a single large group of transcripts specific to the PKC/calcineurin cluster as in [4], we find this group split among other TCs. Hughes et al. identified this group as comprising genes activated when yeast are treated with FK506 or cyclosporin-A. Both compounds affect calcineurin, a serine/threonine phosphatase implicated in intracellular ion homeostasis, adaptation to mating pheromone treatment, and mitosis. However, the two compounds are thought to act through different pathways. Hughes et al. list 42 transcripts as part of this PKC/calcineurin gene cluster. Of these, we find 31 in five different TCs. Ten transcripts are found in TC 2 (cell wall), 11 in TC 12 (cell wall), eight in TC 16 (beta-alanine biosynthesis), and

TABLE 2
List of Mutants in Experimental Conditions Clusters (ECs)

<p>Cluster 1 ssn6 tup1</p> <p>Cluster 2 CDC42 (TET promoter) KAR2 (TET promoter)</p> <p>Cluster 3 HU rad6 rnr1 swi6</p> <p>Cluster 4 rpd3 sin3</p> <p>Cluster 5 dig1 dig1, dig2 dig1-, dig2- (haploid) fus3- (haploid) hda1 hog1- (haploid) sst2- (haploid) yor080w</p> <p>Cluster 6 dot4 mrt4 rpl27a- rps24a- rps24a- (haploid) rps27b- rrp6 sir4 yel033w yel044w yhr034c ymr014w ymr269w</p> <p>Cluster 7 ca884-vs.-ca881 Calcofluorwhite ERG11- (TET promoter) erg2 erg3- (haploid) hmg1- (haploid) HMG2 (TET promoter) imp2' Itraconazole Lovastatin sir2 Terbinafine top3- (haploid) yer044c</p>	<p>Cluster 8 aj307-vs.-aj308 ca721-vs.-ca702 fus3-, kss1- (haploid) med2- (haploid) sgt2 sod1- (haploid) ste11- (haploid) ste12- (haploid) ste12- (haploid) ste18- (haploid) ste4- (haploid) ste5- (haploid) ste7- (haploid) yjl107c- (haploid)</p> <p>Cluster 9 cup5 qcr2- (haploid) rip1 vma8</p> <p>Cluster 10 Tunicamycin yer083c</p> <p>Cluster 11 ade2 aj307-vs.-aj308 bim1 bub1 bub3 bul1 cka2 erg4 pfd2 rtg1 rts1 vac8 vps8</p> <p>Cluster 12 isw1 isw1, isw2 isw2 ras2- (haploid)</p>	<p>Cluster 13/14 2-deoxy-D-glucose anp1- AUR1 (TET promoter) clb2 CMD1 (TET promoter) erg4- (haploid) fks1- (haploid) FKS1 (TET promoter) gas1 Glucosamine hst3 kin3 rad57 she4 spfl swi4 swi5 yar014c</p> <p>Cluster 15 aep2 afg3- (haploid) ard1 ase1 bub3- (haploid) ca719-vs.-ca700 cem1 cyt1 imp2 kim4 mac1 mrp133 msu1 rml2- yap1 yer050c yhl029c yhr011w- ymr031w-a ymr293c</p> <p>Cluster 16 aj318-vs.-aj317 aj324-vs.-aj323 aj338-vs.-aj337 arg80 ca1047-vs.-ca1048 ca1081-vs.-ca1082 ca1083-vs.-ca1084</p>	<p>Cluster 16 (continued) ca1105-vs.-ca1106 ca1107-vs.-ca1108 ca1109-vs.-ca1110 ca1133-vs.-ca1134 ca1135-vs.-ca1136 ca1167-vs.-ca1168 ca1169-vs.-ca1170 ca1171-vs.-ca1172 ca1189-vs.-ca1190 ca1191-vs.-ca1192 ca1290-vs.-ca1289 ca1296-vs.-ca1295 ca1332-vs.-ca1331 ca1334-vs.-ca1333 ca1369-vs.-ca1368 ca1408-vs.-ca1407 ca1410-vs.-ca1409 ca1448-vs.-ca1447 ca1450-vs.-ca1449 ca1488-vs.-ca1487 ca1490-vs.-ca1489 ca1492-vs.-ca1491 ca1547-vs.-ca1546 ca1549-vs.-ca1548 ca1601-vs.-ca1600 ca753-vs.-ca752 ca755-vs.-ca754 ca775-vs.-ca774 ca789-vs.-ca788 ca791-vs.-ca790 ca827-vs.-ca826 ca841-vs.-ca840 ca843-vs.-ca842 ca926-vs.-ca927 ca931-vs.-ca930 ca994-vs.-ca993 cs1412vs.-ca1411 ds1242-vs.-ds1241 ds1244-vs.-ds1243 ds1286-vs.-ds1285 ds1288-vs.-ds1287 ds1308-vs.-ds1307 ds1316-vs.-ds1315 ds720-vs.-ds719 ds798-vs.-ds797 ds800-vs.-ds799 ds866-vs.-ds865 ds904-vs.-ds903 ds906-vs.-ds905 ecm10 gln2 npr2 nta1 pex12 ppr1 sir3</p>
--	---	---	--

one each in TC 4 (monosaccharide transport) and TC 5. It is known that PKC is part of a MAPK cascade involved in cell wall integrity. It has cross talk with other MAPK cascades including pheromone response, osmolarity, and filamentous growth. All told, five of the 17 TCs are associated with the cell wall. Recent work indicates that beyond providing structural support, components of the cell wall are involved in diverse functions from uptake of nutrients/metabolism to energy generation [28]. Likewise, formation of the shmoo during mating involves not only signal transduction by mating factor but also rearrangement of the cytoskeleton and cell wall.

Finally, we identified a cluster (TC 9) that does not appear to be covered by those defined by Hughes et al. The best GO term matches are “siderophore transport” (process

GO), “iron ion transporter activity” (function GO), and “endosome” (component GO).

3.2 Clustering by Experimental Conditions

Clustering of the expression profiles by experimental conditions identifies those yeast mutants or compounds that have similar effects on all transcripts. In Fig. 4, we show that a minimum of about 15 components is necessary, and the data supports up to about 30.

Fig. 5 shows the clustering of the experimental conditions, which has an interpretation similar to that of Fig. 3. After prefiltering the 300 compendium experiments, 194 expression profiles including 60 “wild types” remained. “Wild types” represent control experiments testing neither chemical treatment nor gene knockout, but yet had at least one ORF whose expression changed more than twofold.

TABLE 3
Summary of SGD GO Annotations for TCs

Cluster	#ORFs	Function	p	Process	p	Component	p
1	14	isocitrate dehydrogenase (NAD+) activity	6.83E-06	glutamate biosynthesis	2.03E-06	mitochondrial nucleoid	8.800E-04
2	28	structural constituent of cell wall	5.78E-07	cell wall organization and biogenesis	1.79E-05	cell wall	8.140E-12
3	22	RNA-directed DNA polymerase activity	1.02E-27	Ty element transposition	2.23E-28	retrotransposon nucleocapsid	1.860E-28
4	21	fructose transporter activity	8.60E-04	monosaccharide transport	2.27E-05	plasma membrane	3.144E-02
5	34	molecular function unknown	3.85E-08	biological process unknown	2.40E-07	cellular component unknown	1.170E-06
6	38	cell adhesion molecule binding	5.81E-07	conjugation	2.14E-23	mating projection tip	1.750E-08
7	82	catalytic activity	1.78E-06	amino acid biosynthesis	1.74E-31	carbamoyl-phosphate synthase complex	2.400E-04
8	91	sugar transporter activity	2.79E-11	carbohydrate transport	3.42E-10	cellular component unknown	3.550E-06
9	19	iron ion transporter activity	8.93E-11	siderophore transport	4.02E-19	endosome	3.100E-05
10	10	cyclin-dependent protein kinase regulator activity	4.40E-04	regulation of cyclin dependent protein kinase activity	1.60E-04	endoplasmic reticulum	7.396E-02
11	32	hydrolase activity, hydrolyzing O-glycosyl compounds	4.53E-06	cytokinesis, completion of separation	1.61E-08	cell wall (sensu Fungi)	2.510E-08
12	14	endopeptidase activity	6.14E-03	cell wall organization and biogenesis	3.19E-02	plasma membrane	7.754E-02
13	18	monoxygenase activity	7.14E-07	steroid biosynthesis	6.23E-19	endoplasmic reticulum	4.400E-11
14	8	oxidoreductase activity, acting on sulfur group of donors	2.48E-07	sulfur utilization	2.71E-13	sulfite reductase complex (NADPH)	2.100E-06
15	17	oxidoreductase activity, acting on the aldehyde or oxo group of donors	1.54E-03	vitamin metabolism	1.40E-02	storage vacuole	4.414E-02
16	42	protease inhibitor activity	1.86E-06	beta-alanine biosynthesis	6.43E-05	cytoplasm	1.090E-03
17	25	polyamine transporter activity	1.57E-03	polyamine transport	5.50E-04	vacuole	3.830E-05

TABLE 4
Summary of SGD GO Annotations for ECs

Cluster	#ORFs	Function	p	Process	p	Component	p
1	2	general transcriptional repressor activity	7.52E-08	nucleosome spacing	1.69E-07	nucleus	7.40E-02
2	2	nucleoside-triphosphatase activity	1.03E-03	conjugation	2.00E-04	none	-
3	3	-	-	nucleobase, nucleoside, nucleotide and nucleic acid metabolism	9.15E-03	protein complex	9.97E-02
4	2	histone deacetylase activity	5.43E-06	chromatin silencing at rDNA	3.68E-06	histone deacetylase complex	7.52E-06
5	7	MAP kinase activity	9.85E-06	invasive growth (sensu Saccharomyces)	1.63E-08	nucleus	1.88E-02
6	12	structural molecule activity	2.11E-03	rRNA processing	2.42E-03	non-membrane-bound organelle	3.90E-04
7	8	hydroxymethylglutaryl-CoA reductase (NADPH) activity	2.10E-06	ergosterol metabolism	4.52E-14	endoplasmic reticulum	2.38E-07
8	12	receptor signaling protein activity	8.75E-09	invasive growth (sensu Saccharomyces)	3.72E-14	mating projection	5.76E-07
9	5	hydrogen ion transporter activity	4.01E-06	hydrogen ion homeostasis	9.88E-05	hydrogen-translocating V-type ATPase complex	4.21E-05
10	-	-	-	-	-	-	-
11	12	protein binding	7.70E-04	spindle checkpoint	4.45E-06	kinetochore	8.50E-05
12	4	nucleoside-triphosphatase activity	1.20E-04	chromatin remodeling	1.94E-03	chromatin remodeling complex	5.90E-04
13	10	transferase activity, transferring hexosyl groups	4.74E-03	protein amino acid glycosylation	3.83E-03	incipient bud site	1.01E-03
14	-	-	-	protein amino acid glycosylation	3.08E-03	-	-
15	23	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides	5.75E-03	physiological process	6.27E-05	mitochondrion	2.23E-05
16	41	-	-	wild type	-	-	-

- : no value determined.

TABLE 5
Summary of SGD GO Annotations for ECs Described by Hughes et al. [4]

Hughes assignment	#ORFs	Function	p	Process	p	Component	p
1 mitochondrial function	17	oxidoreductase activity, acting on diphenols and related substances as donors, cytochrome as acceptor	2.00E-04	mitochondrial electron transport, ubiquinol to cytochrome c	1.26E-06	mitochondrion	2.38E-11
2 cell wall	9	catalytic activity	2.05E-03	budding cell apical bud growth	8.13E-05	actin cap	1.50E-04
3 protein synthesis	12	structural constituent of ribosome	4.00E-04	ribosome biogenesis and assembly	5.40E-04	intracellular non-membrane-bound organelle	3.80E-05
4 ergosterol biosynthesis	8	hydroxymethylglutaryl-CoA reductase (NADPH) activity	1.12E-06	ergosterol metabolism	1.62E-15	endoplasmic reticulum	9.24E-09
5 mating	8	receptor signaling protein activity	1.24E-09	invasive growth (sensu Saccharomyces)	3.82E-16	cell projection	8.31E-08
6 MAPK activation	6	MAP kinase activity	7.04E-06	filamentous growth	3.56E-07	plasma membrane	1.55E-02
7 rnr1/HU	3	none	-	nucleobase, nucleoside, nucleotide and nucleic acid metabolism	9.15E-03	protein complex	9.97E-02
8 histone deacetylase	3	histone deacetylase activity	1.26E-08	chromatin silencing at rDNA	7.07E-09	histone deacetylase complex	2.06E-08
9 isw	3	ATPase activity	1.68E-03	chromatin remodeling	6.07E-06	chromatin remodeling complex	2.90E-04
10 vacuolar ATPase/iron regulation	3	hydrogen ion transporter activity	1.60E-04	cation homeostasis	1.94E-06	hydrogen-translocating V-type ATPase complex	1.26E-05
11 sir	3	histone binding	1.65E-10	loss of chromatin silencing during replicative cell aging	1.32E-09	nuclear telomeric heterochromatin	8.84E-10
12 tup1/ssn6	2	general transcriptional repressor activity	7.52E-08	nucleosome spacing	1.69E-07	nucleus	7.40E-02

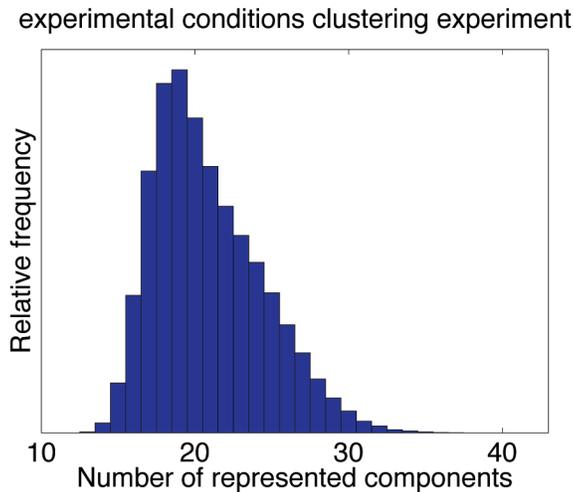


Fig. 4. Number of represented components for the clustering of the experimental conditions. This figure shows that the relative number of components along the MCMC run varies between about 15 and 30.

(These were explicitly excluded from the cluster analysis of Hughes et al.)

From Fig. 5, 16 experiment clusters (ECs) are apparent. This is in contrast to the 13 identified by Hughes et al. [4]. As with the transcript response clustering, it can be seen that some clusters are bipartite (e.g., ECs 5, 7, and 11), and there is a region of diffuse clusters (ECs 9-14). Closer examination suggests there may be smaller clusters within this region. Also, two clusters (ECs 13 and 14) may be considered to be overlapping. In addition, a dendrogram using the dissimilarity measure defined above is shown in Fig. 6, which may be compared to the supplementary material of [4, Fig. 3B].

Apart from EC 1, other ECs correspond closely, although not exactly, to those identified by Hughes et al. For example, the Hughes et al. cluster *rrr1/HU* overlaps with our EC 3 with the exception of MMS. We both find a histone deacetylase group (EC 4), an ergosterol biosynthesis group (EC 7), a mating group (EC 8), a V-ATPase/iron regulation group (EC 9), and a mitochondrial group (EC 15). The "ribosome/translation" group identified by Hughes et al. overlaps with EC 6, which is associated with the molecular process GO term of "rRNA processing."

A major difference between Hughes et al. and our DPM results involves profiles identified as "cell wall." Hughes et al. identified 13 expression profiles as part of a "cell wall" group. However, we find three distinct clusters within this group. Knockouts for two tetracycline-driven genes, *tet-KAR2* and *tet-CDC42*, cluster together as EC 2 with a co-occurrence probability close to 100 percent; this cluster does not overlap with any other. In addition, tunicamycin and yer083 form a cluster (identified as EC 10) with a co-occurrence probability around 85 percent, clearly apart from other profiles. Tunicamycin is thought to disrupt protein glycosylation in yeast [29], while yer083c has recently been identified as localized to the ER and involved in trafficking cell wall proteins [30], [31]. The remaining members appear in EC 13 which is associated with "incipient bud site" as its best component GO term. Thus, while all 13 members do

involve proteins associated with the cell wall, it may be seen that multiple processes or functions are being affected. Recent work has indicated the cell wall stress influences many genes through diverse signaling pathways and different transcription factors [27], [32].

Hughes et al. identify a single cluster containing the *sir* mutants. Sir proteins are involved in global gene regulation through chromatin restructuring. However, by DPM clustering, we find each *sir* knockout in a different cluster: *sir2Δ* in EC 7 (ergosterol), *sir3Δ* in EC 16 (wild type), and *sir4Δ* in EC 6 (rRNA processing). We note that association of *sir2* with EC 7 is at a co-occurrence probability of 60 percent, and association of *sir4Δ* with EC 6 is at 30 percent. This suggests that while the SIR proteins are not strongly affiliated with any other group or each other globally, there may be a subset of specific transcripts that are strongly affected. It is possible that while there are few coregulated transcripts, their regulation may be highly similar. The expression profile of the *sir2Δ* mutant is most similar to that of *imp2'* (YIL154C) at a co-occurrence probability close to 80 percent. Sir2p is involved in chromatin silencing; disruption causes problems with DNA repair while slight overexpression increases the life span of yeast and *C. elegans* [33], [34]. It is known that caloric restriction increases Sir2p activity. Imp2p is a transcription factor that activates galactose, maltose, and raffinose utilization [35] as well as mediating oxidative damage to DNA [36]. Similarity in the expression profiles of these two mutants might be because the set of genes derepressed by the *sir2Δ* mutant overlap somewhat with those regulated by Imp2p. Alternately, both mutants might exhibit similar global effects.

The *isw1*, *isw2* group found by Hughes et al. contains four expression profiles (*isw1*, *isw2*, *isw1/2*, and *hst3*). We identify EC 12 containing *isw1*, *isw1/isw2*, *isw2*, and *ras2* but instead put *hst3* as part of the larger ECs 13/14. The ISW proteins are ATPases and are likely part of a protein complex involved in chromatin remodeling [37]. Ras2p is a GTP-binding protein involved in nitrogen starvation response, sporulation, and filamentous growth [38]. Hst3p is part of the Sir protein family of histone deacetylases and thought to be involved multiple functions including telomeric silencing [39]. As noted above, while Hughes et al. placed the Sir proteins into a single cluster, we find them distributed through several clusters. However, examination of the dendrogram (Fig. 6) indicates that EC 12 may be considered a "subcluster" within the larger ECs 13/14 and is joined to the subcluster containing *hst3Δ*.

3.3 Discussion

Although the use of clustering methods (in particular, agglomerative hierarchical clustering) has rapidly become one of the standard computational approaches in the literature of microarray gene expression data, little attention has been paid to *uncertainty* in the results obtained.

Kerr and Churchill [40] have proposed the use of a bootstrap method to assess the results of clustering in a statistically quantifiable manner. However, their approach requires the fitting of a linear statistical (ANOVA) model to the microarray data to obtain least squares estimates of the differential expression of a given gene, which are then used

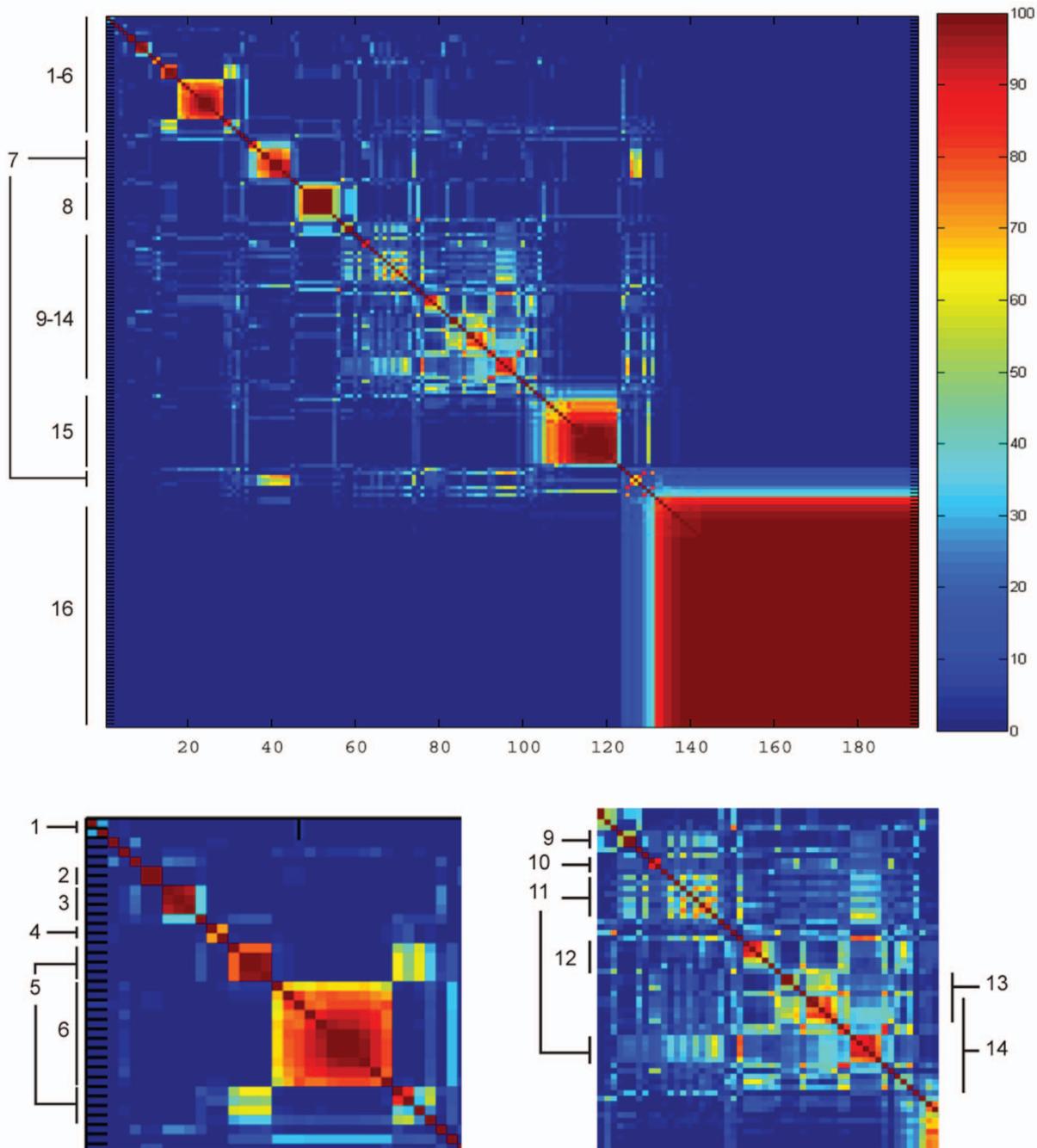


Fig. 5. Co-occurrence probabilities of the 194 ECs. This figure shows the number of times, out of 100 samples, that the *indicator variables* for two experimental conditions were equal. This may be interpreted as the probability p_{ij} that two experimental conditions i and j belong to the same cluster, and the different colors represent this probability. Numbers 1-16 indicated in the margins refer to the ECs discussed in detail in the text. The subfigures represent a magnified view of portions of the larger figure. A larger version of this figure is available in the supplementary materials, which can be found on the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TCBB.2007.70269>.

as inputs to the bootstrap process. An alternative parametric bootstrap approach has been described by Zhang and Zhao [41] which uses estimates of the standard errors in gene expression measurements to simulate data from a lognormal distribution. Hughes et al. [4] describe a permutation procedure to calculate p -values for the significance of branching in a dendrogram produced by agglomerative hierarchical clustering, under the null hypothesis that the branching was not significant. However, hierarchical clustering is a bottom-up algorithm. It starts

with each data point assigned to its own cluster and iteratively merges the two closest clusters together until all the data belongs to a single cluster. Consequently, the results presented by Hughes et al. (supplementary information to [4, Fig. 3B]) only appear to show strong confidence for the branches at the lowest level of the dendrogram. In contrast, the dendrogram produced from the DPM procedure (Fig. 6) represents a full probabilistic measure of the (dis)similarity of two gene expression profiles.

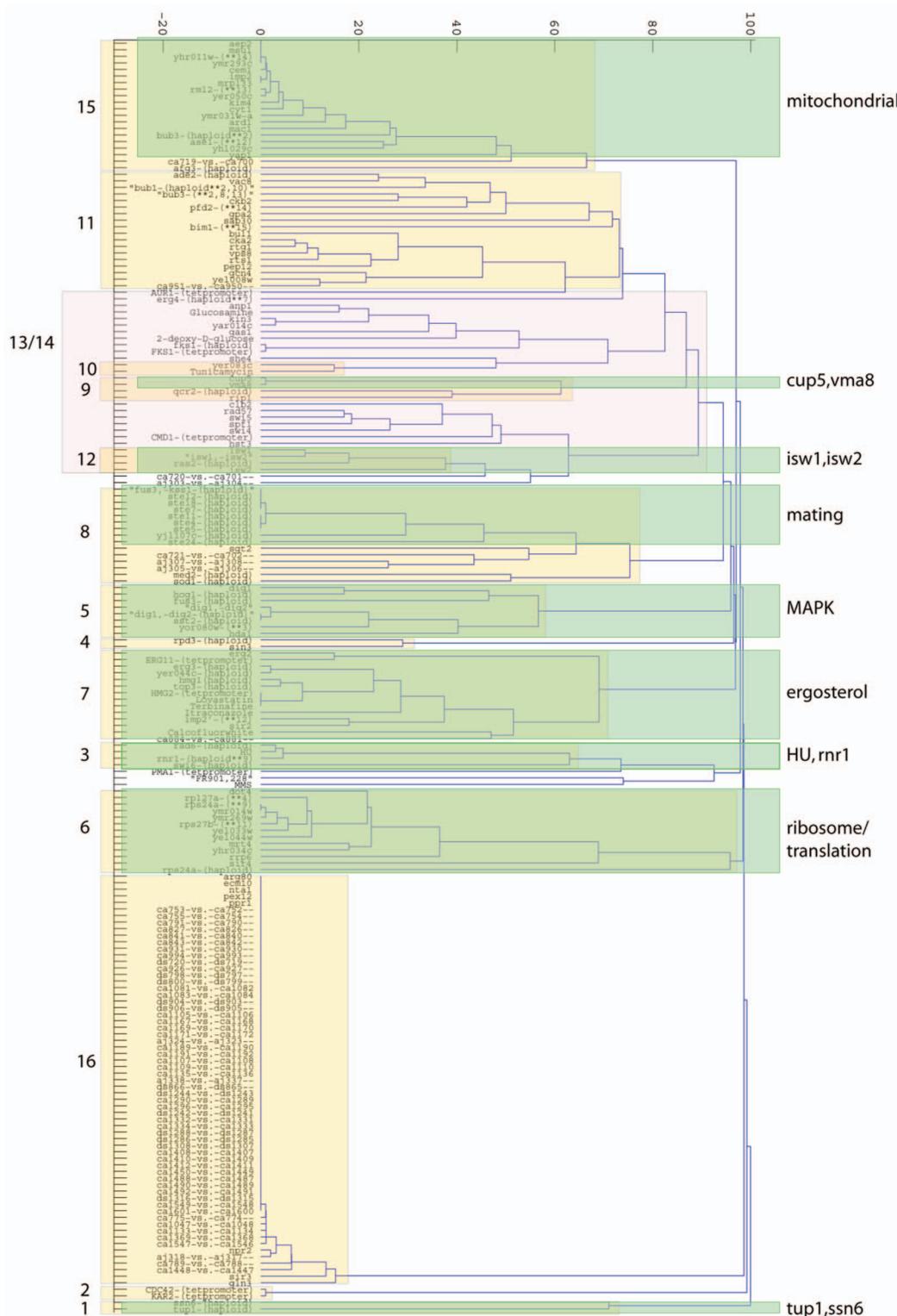


Fig. 6. Dendrogram of the dissimilarity measure of the 194 ECs. $1 - D_{ij}$ defines a dissimilarity measure, which can be input to one of the standard linkage algorithms used for hierarchical clustering. Results are shown for the dendrogram obtained by average linkage. The numbers on the left side of the figure and yellow and pink boxes indicate ECs discussed in detail in the text. Labels on the right-hand side (and green boxes) indicate clusters assigned by Hughes et al. A larger version of this figure is available in the supplementary materials, which can be found on the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TCBB.2007.70269>.

DPM models provide a nonparametric Bayesian alternative to the bootstrap approach to modeling uncertainty in gene expression clustering. Medvedovic et al. have applied infinite Gaussian mixture (or DPM) models to the clustering

of time series gene expression data using spherical Gaussians with diagonal covariances [12], [13]. Similar approaches have also recently been described in [16]. However, these approaches do not explicitly model the

correlations between subsequent time points which would be expected to occur in time series data, and the use of diagonal covariances may result in more clusters than necessary to model such correlations. Lui et al. have recently extended their previous work to use full-covariance models for time series [14]. Since these authors are clustering short time series, inference in the space of the original data is feasible. In contrast, in the complementary approach we describe here, we apply the DPM method to high-dimensional nontime series data. Inference is carried out in a low-dimensional projection of this space after dimensionality reduction by principal component analysis, which makes it possible to use Gaussians with full-covariance matrices, which would be very computationally expensive in the original high-dimensional space as each sampling step has a cubic computational dependency on the dimensionality.

Bayesian approaches to clustering gene expression data have recently received much attention. Heard et al. [42] propose an agglomerative clustering procedure for gene expression time series curves based on a Bayesian merging score, but unrelated to DPMs. Heller and Ghahramani [43] proposed a different Bayesian hierarchical clustering (BHC) procedure which implements a non-MCMC inference procedure for DPMs. This BHC algorithm can be used to scale DPM learning and inference to very large data sets at the cost of only partially representing the uncertainty in the cluster assignments. The MCMC procedure we present in this paper is more computationally demanding but captures more completely the sources of uncertainty. In [44], model-based clustering procedures based on loss functions are derived. An integer program is identified for finding a single clustering that best matches the posterior co-occurrence probabilities.

Recently, Bidaut et al. [45] have reanalyzed the data of Hughes et al. using “Bayesian decomposition” to place the experimental profiles into patterns (clusters). The highest scoring (high persistence) genes in the patterns were annotated using the MIPS database [46] to assign the pattern to a cellular pathway. Fifteen patterns were discovered, six of which are assigned to MIPS pathways. Bidaut et al. find that *ssn6Δ* and *tup1Δ* appear in many of their patterns, albeit at low persistence. In contrast, with DPM modeling, we find that *ssn6Δ* and *tup1Δ* cluster together although weakly (EC 1—co-occurrence probability of 30 percent) and apart from other experimental profiles. This is reinforced by the dendrogram (Fig. 6) which shows while the *tup1Δ* and *ssn6Δ* profiles cluster away from the others, they are yet on very long branches from each other. Clustering of these two knockouts is supported by the fact that Tup1p and Ssn6p are thought to form a protein complex. As previously mentioned, both proteins are transcription factors involved in glucose/catabolic repression although with different but overlapping sets of targets [39].

Patterns 13 and 15 identified by Bidaut et al. [45] are given significance as distinguishing between those genes involved in MAPK signaling mating versus those involved in filamentous growth. While these are two distinct cellular functions, they share signaling components. Bidaut et al. suggest that these groups can be distinguished by whether the genes are regulated by Ste12p or the Ste12p-Tec1p complex. In our clustering of experimental conditions, all of the *ste* deletion mutants plus the *fus3Δ*, *kss1Δ* double

mutant cluster together (EC 8—component GO term: mating projection). The *fus3Δ* single mutant appears in EC 8, together with other genes annotated by the GO molecular function term indicating MAPK activity. However, when we look at the top genes associated with the Bidaut patterns, six of the top 10 genes in pattern 13 are part of TC 6 (component GO: mating projection tip), while seven of the top 10 genes in pattern 15 are part of our TC 3 (component GO: retrotransposon nucleocapsid).

4 CONCLUSION

DPM models provide a nonparametric Bayesian alternative to the bootstrap approach to modeling uncertainty in gene expression clustering. Unlike methods based on a single clustering of the data, the approach computes the probability that two genes belong to the same cluster while taking into account the main sources of model uncertainty, including the number of clusters and the location of clusters. Biologically plausible results are obtained from the Rosetta compendium of expression profiles which extend previously published cluster analyses of this data. Our results not only confirm many of the previously published clusters identified in this data set but also provide new biological insights by revealing a finer level of granularity in the clustering. In particular, our method was able to distribute general stress response and carbohydrate metabolism and amino acid biosynthesis groups into more specific clusters. While previous analyses have identified a group of profiles collectively related to cell wall functions, our results also suggest that this large group forms smaller, distinctly regulated subclusters. These results are consistent with recent literature on cell wall proteins which suggests that distinct functions—for example, controlling osmotic pressure, responding to physical stress, maintaining cell wall integrity, and providing a protein scaffold—may share proteins and have overlapping regulatory mechanisms.

APPENDIX

Additional files,

- File 1—Tab-delimited file of Component GO terms for Transcript Response Clusters: Supp_Table_1.txt,
 - File 2—Tab-delimited file of Functional GO terms for Transcript Response Clusters: Supp_Table_2.txt,
 - File 3—Tab-delimited file of Process GO terms for Transcript Response Clusters: Supp_Table_3.txt,
 - File 4—Tab-delimited file of Component GO terms for Experimental Clusters: Supp_Table_4.txt,
 - File 5—Tab-delimited file of Functional GO terms for Experimental Clusters: Supp_Table_5.txt,
 - File 6—Tab-delimited file of Process GO terms for Experimental Clusters: Supp_Table_6.txt,
 - File 7—Large version of Figure 3: Fig03-largeNew.tif,
 - File 8—Large version of Figure 5: Fig05-largeNew.tif, and
 - File 9—Large version of Figure 6: Fig06-largeNew.tif,
- can be found on the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TCBB.2007.70269>.

REFERENCES

- [1] M. Eisen, P. Spellman, P. Brown, and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression," *Proc. Nat'l Academy of Sciences USA*, vol. 95, pp. 14863-14868, 1998.
- [2] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proc. Nat'l Academy of Sciences USA*, vol. 96, pp. 6745-6750, 1999.
- [3] G. McLachlan, R. Bean, and D. Peel, "A Mixture Model-Based Approach to the Clustering of Microarray Expression Data," *Bioinformatics*, vol. 18, no. 3, pp. 413-422, 2002.
- [4] T. Hughes, M. Marton, A. Jones, C. Roberts, R. Stoughton, C. Armour, H. Bennett, E. Coffey, H. Dai, Y. He, M. Kidd, A. King, M. Meyer, D. Slade, P. Lum, S. Stepaniants, D. Shoemaker, D. Gachotte, K. Chakraburttty, J. Simon, M. Bard, and S. Friend, "Functional Discovery via a Compendium of Expression Profiles," *Cell*, vol. 102, pp. 109-126, July 2000.
- [5] R.M. Neal, "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *J. Computational and Graphical Statistics*, vol. 9, pp. 249-265, 2000.
- [6] C.E. Rasmussen, "The Infinite Gaussian Mixture Model," *Advances in Neural Information Processing Systems 12*, S.A. Solla, T.K. Leen, and K.-R. Müller, eds., pp. 554-560, MIT Press, 2000.
- [7] C. Antoniak, "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," *Annals of Statistics*, vol. 2, pp. 1152-1174, 1974.
- [8] T. Ferguson, "A Bayesian Analysis of Some Nonparametric Problems," *Annals of Statistics*, vol. 1, pp. 209-230, 1973.
- [9] A.Y. Lo, "On a Class of Bayesian Nonparametric Estimates: I. Density Estimates," *Annals of Statistics*, vol. 12, pp. 351-357, 1984.
- [10] M.D. Escobar and M. West, "Bayesian Density Estimation and Inference Using Mixtures," *J. Am. Statistical Assoc.*, vol. 90, no. 430, pp. 577-588, 1995.
- [11] D.L. Wild, C.E. Rasmussen, Z. Ghahramani, J. Cregg, B.J. de la Cruz, C.-C. Kan, and K.A. Scanlon, "A Bayesian Approach to Modelling Uncertainty in Gene Expression Clusters," *Proc. Third Int'l Conf. Systems Biology (ICSB)*, 2002.
- [12] M. Medvedovic and S. Sivaganesan, "Bayesian Infinite Mixture Model Based Clustering of Gene Expression Profiles," *Bioinformatics*, vol. 18, no. 9, pp. 1194-1206, 2002.
- [13] M. Medvedovic, K.Y. Yeung, and R.E. Bumgarner, "Bayesian Mixture Model Based Clustering of Replicated Microarray Data," *Bioinformatics*, vol. 20, no. 8, pp. 1222-1232, 2004.
- [14] X. Liu, S. Sivaganesan, K.Y. Yeung, J. Guo, R.E. Bumgarner, and M. Medvedovic, "Context-Specific Infinite Mixtures for Clustering Gene Expression Profiles across Diverse Microarray Dataset," *Bioinformatics*, vol. 22, no. 14, pp. 1737-1744, 2006.
- [15] D. Dahl, "Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model," *Bayesian Inference for Gene Expression and Proteomics*, K.-A. Do, P. Müller, and M. Vannucci, eds., Cambridge Univ. Press, 2006.
- [16] Z.S. Qin, "Clustering Microarray Gene Expression Data Using Weighted Chinese Restaurant Process," *Bioinformatics*, vol. 22, no. 16, pp. 1988-1997, 2006.
- [17] A. Dubey, S. Hwang, C. Rangel, C. Rasmussen, Z. Ghahramani, and D.L. Wild, "Clustering Protein Sequence and Structure Space with Infinite Gaussian Mixture Models," *Proc. Pacific Symp. Biocomputing (PSB '04)*, R.B. Altman, A.K. Dunker, L. Hunter, and T.E. Klein, eds., pp. 399-410, 2004.
- [18] J. Hartigan, *Clustering Algorithms*. Wiley, 1975.
- [19] K. Yeung, D. Haynor, and W. Ruzzo, "Validating Clustering for Gene Expression Data," *Bioinformatics*, vol. 17, pp. 309-318, 2001.
- [20] D.J. Mackay, *Information Theory, Inference and Learning Algorithms*. Cambridge Univ. Press, 2003.
- [21] G. McLachlan and D. Peel, *Finite Mixture Models*. Wiley, 2000.
- [22] K. Yeung, C. Fraley, A. Murua, A. Raftery, and W. Ruzzo, "Model Based Clustering and Data Transformations for Gene Expression Data," *Bioinformatics*, vol. 17, pp. 977-987, 2001.
- [23] D. Görür, "Nonparametric Bayesian Discrete Latent Variable Models for Unsupervised Learning," PhD dissertation, Max Planck Inst. for Biological Cybernetics, 2007.
- [24] E. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. Cherry, and G. Sherlock, "Go::Termfinder-Open Source Software for Accessing Gene Ontology Information and Finding Significantly Enriched Gene Ontology Terms Associated with a List of Genes," *Bioinformatics*, vol. 20, no. 18, pp. 3710-3715, 2004.
- [25] M. Viswanathan, G. Muthukumar, Y.S. Cong, and J. Lenard, "Seripauperins of Saccharomyces Cerevisiae: A New Multigene Family Encoding Serine-Poor Relatives of Serine-Rich Proteins," *Gene*, vol. 148, no. 1, pp. 149-153, 1994.
- [26] N. Rachidi, M.J. Martinez, P. Barre, and B. Blondin, "Saccharomyces Cerevisiae PAU Genes Are Induced by Anaerobiosis," *Molecular Microbiology*, vol. 35, no. 6, pp. 1421-1430, 2000.
- [27] F. Klis, A. Boorsma, and P.D. Groot, "Cell Wall Construction in Saccharomyces Cerevisiae," *Yeast*, vol. 23, no. 185-202, 2006.
- [28] U. Jung and D. Levin, "Genome-Wide Analysis of Gene Expression Regulated by the Yeast Cell Wall Integrity Signalling Pathway," *Molecular Microbiology*, vol. 34, pp. 1049-1057, 1999.
- [29] W. McDowell and R. Schwarz, "Dissecting Glycoprotein Biosynthesis by Use of Specific Inhibitors," *Biochimie*, vol. 70, pp. 1535-1549, 1998.
- [30] A. Enyenihi and W. Saunders, "Large-Scale Functional Genomic Analysis of Sporulation and Meiosis in Saccharomyces Cerevisiae," *Genetics*, vol. 163, no. 1, pp. 47-54, 2003.
- [31] M. Schuldiner et al., "Exploration of the Function and Organization of the Yeast Early Secretory Pathway through an Epistatic Miniarray Profile," *Cell*, vol. 123, no. 3, pp. 507-519, 2005.
- [32] A. Boorsma, H. de Nobel, B. ter Riet, B. Bargmann, S. Brul, K. Hellingwerf, and F. Klis, "Characterization of the Transcriptional Response to Cell Wall Stress in Saccharomyces Cerevisiae," *Yeast*, vol. 21, pp. 413-427, 2004.
- [33] M. Kaeberlein, M. McVey, and L. Guarente, "The sir2/3/4 Complex and sir2 Alone Promote Longevity in Saccharomyces Cerevisiae by Two Different Mechanisms," *Genes and Development*, vol. 13, pp. 2570-2580, 1999.
- [34] G. Blander and L. Guarente, "The sir2 Family of Protein Deacetylases," *Ann. Rev. Biochemistry*, vol. 73, pp. 417-435, 2004.
- [35] J. Masson and D. Ramotar, "The Saccharomyces Cerevisiae imp2 Gene Encodes a Transcriptional Activator that Mediates Protection against DNA Damage Caused by Bleomycin and Other Oxidants," *Molecular and Cellular Biology*, vol. 16, no. 5, pp. 2091-2100, 1996.
- [36] C. Domini et al., "Imp2, a Nuclear Gene Controlling the Mitochondrial Dependence of Galactose, Maltose and Raffinose Utilization in Saccharomyces Cerevisiae," *Yeast*, vol. 8, no. 2, pp. 83-93, 1992.
- [37] J. Mellor and A. Morillon, "Iswi Complexes in Saccharomyces Cerevisiae," *Biochimica et Biophysica Acta*, vol. 1677, nos. 1-3, pp. 100-112, 2004.
- [38] T. Kataoka et al., "Genetic Analysis of Yeast ras1 and ras2 Genes," *Cell*, vol. 37, no. 2, pp. 437-445, 1984.
- [39] R.L. Smith and A.D. Johnson, "Turning Genes Off by ssn6-tup1: A Conserved System of Transcriptional Repression in Eukaryotes," *Trends in Biochemical Sciences*, vol. 25, no. 325-330, 2000.
- [40] M.K. Kerr and G.A. Churchill, "Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments," *Proc. Nat'l Academy of Sciences USA*, vol. 98, no. 16, pp. 8961-8965, 2001.
- [41] K. Zhang and H. Zhao, "Assessing Reliability of Gene Clusters from Gene Expression Data," *Functional & Integrative Genomics*, vol. 1, pp. 156-173, 2000.
- [42] N.A. Heard, C.C. Holmes, and D.A. Stephens, "A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves," *J. Am. Statistical Assoc.*, vol. 101, no. 473, pp. 18-29, 2006.
- [43] K.A. Heller and Z. Ghahramani, "Bayesian Hierarchical Clustering," *Proc. 22nd Int'l Conf. Machine Learning (ICML)*, 2005.
- [44] J.W. Lau and P.J. Green, "Bayesian Model Based Clustering Procedures," *J. Computational and Graphical Statistics*, vol. 16, no. 3, pp. 526-558, 2007.
- [45] G. Bidaut, K. Suhre, J.-M. Claverie, and M. Ochs, "Determination of Strongly Overlapping Signaling Activity from Microarray Data," *BMC Bioinformatics*, vol. 7, pp. 99-111, 2006.
- [46] H. Mewes et al., "Mips: Analysis and Annotation of Proteins from Whole Genomes," *Nucleic Acids Research*, vol. 32, pp. D41-D44, 2004.



Carl Edward Rasmussen received the master's degree in engineering from the Technical University of Denmark and the PhD degree in computer science from the University of Toronto, Canada, in 1996. Since then, he has been a postdoctoral researcher at the Technical University of Denmark, a senior research fellow in the Gatsby Computational Neuroscience Unit, University College London from 2000 to 2002, and a junior research group leader at the Max

Planck Institute for Biological Cybernetics, Tübingen, Germany, from 2002 to 2007. He is a lecturer in the Computational and Biological Learning Lab, Department of Engineering, University of Cambridge, and an adjunct research scientist at the Max Planck Institute for Biological Cybernetics, Tübingen. His main research interests include Bayesian inference and machine learning.



Zoubin Ghahramani received the BA and BSE degrees from the University of Pennsylvania and the PhD degree from MIT, in 1995, working with Professor Mike Jordan. He is a professor of information engineering at the University of Cambridge, Cambridge, United Kingdom, and is also an associate research professor of machine learning at Carnegie Mellon University, Pittsburgh. He did his postdoctoral work in computer science at the University of Toronto,

working with Professor Geoff Hinton. He was a faculty member in the Gatsby Unit, University College London from 1998 to 2005. His work has included research on human sensorimotor control, cognitive science, statistics, and machine learning. His current interests include Bayesian approaches to statistical machine learning, nonparametric methods, graphical models, and approximate inference. He is also actively working on applications of machine learning to bioinformatics and information retrieval. He has published more than 100 peer-reviewed papers. He serves on the editorial boards of several leading journals in the field, including the *Journal of Machine Learning Research*, *Annals of Statistics*, the *Journal of Artificial Intelligence Research*, *Machine Learning*, *Foundations and Trends in Machine Learning*, and *Bayesian Analysis*. He is an associate editor-in-chief of the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, serves on the board of the International Machine Learning Society, and was the program chair of the 2007 International Conference on Machine Learning.



Bernard J. de la Cruz received the BA degree in biology from the University of California, Santa Cruz, and the PhD degree in biology from the University of California, San Diego. He was a postdoctoral research fellow at the Keck Graduate Institute, studying the metabolism and genome of the yeast *Pichia pastoris*. His research interests include metabolic regulation of genes and analysis of microarray data. He is currently in the Beck-

man Research Institute at the City of Hope.



David L. Wild received the BA degree in physics from the University of York and the DPhil degree in molecular biophysics from the University of Oxford. He has extensive experience in structural and computational molecular biology. He has worked in the European Molecular Biology Laboratory, the Salk Institute, and in industry with Allelix Biopharmaceuticals, Oxford Molecular, and GlaxoWellcome. He is currently a professor of bioinformatics at the

University of Warwick Systems Biology Centre and an adjunct research professor at the Keck Graduate Institute of Applied Life Sciences. His research interests include bioinformatics, systems, and structural biology.

▷ **For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.**