Physiol Genomics
4: 127–135, 2000.

Integrating naive Bayes models and external knowledge to examine copper and iron homeostasis in *S. cerevisiae*

E. J. MOLER,* D. C. RADISKY,* AND I. S. MIAN
Department of Cell and Molecular Biology, Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720

Received 4 April 2000; accepted in final form 10 September 2000

Moler, E. J., D. C. Radisky, and I. S. Mian. Integrating naive Bayes models and external knowledge to examine copper and iron homeostasis in *S. cerevisiae. Physiol Genomics* 4: 127–135, 2000.—A novel suite of analytical techniques and visualization tools are applied to 78 published transcription profiling experiments monitoring 5,687 *Saccharomyces cerevisiae* genes in studies examining cell cycle, responses to stress, and diauxic shift. A naive Bayes model discovered and characterized 45 classes of gene profile vectors. An enrichment measure quantified the association between these classes and specific external knowledge defined by four sets of categories to which genes can be assigned: 106 protein functions, 5 stages of the cell cycle, 265 transcription factors, and 16 chromosomal locations. Many of the 38 genes in *class 42* are known to play roles in copper and iron homeostasis. The 17 uncharacterized open reading frames in this class may be involved in similar homeostatic processes; human homologs of two of them could be associated with as yet undefined disease states arising from aberrant metal ion regulation. The Met4, Met31, and Met32 transcription factors may play a role in coregulating genes involved in copper and iron metabolism. Extensions of the simple graphical model used for clustering to learning more complex models of genetic networks are discussed.

molecular profile matrix; gene profile vectors; naive Bayes model; copper and iron metabolism; Bayesian networks

METHODS FOR THE ANALYSIS of transcription profile data obtained from high density oligonucleotide or cDNA microarrays include hierarchical clustering (6), gene shaving (9), self-organizing maps (SOMs) (26, 28), Boolean networks (13, 14, 24), linear modelling (5), principal component analysis (21), nonlinear modeling (29), Bayesian networks (BNs) (7), dynamic Bayesian networks (DBNs) (18), Support Vector Machines (SVMs) (3, 17)), and Petri nets (8, 15). Recently, a modular framework that combines generative and discriminative methods was proposed for the analysis of profile data and domain knowledge with the goal of elucidating basic mechanisms and pathways and developing decision support systems for diagnosis, prognosis, and monitoring (17). A molecular profile matrix was defined as the concatenation of multiple profiling experiments in which each row, a molecule profile vector, is the profile of a molecule under different conditions, and each column, an experiment profile vector, is an individual experiment.

Working prototypes of techniques and tools that address tasks in specific modules were applied to transcription profile data from human colon adenocarcinoma tissue specimens, namely sixty-two 1,988-feature experiment profile vectors labeled tumor or nontumor (2). A naive Bayes model discovered and characterized three classes (clusters) of profile vectors and thus subtypes of the specimens (unsupervised learning). SVMs distinguished tumor from nontumor specimens and assigned the label of profile vectors not used for training (supervised learning). Fifty to 200 genes were identified that distinguished the two types of specimens as well as or better than the 1,988 assayed originally (feature relevance, ranking, and selection). This small subset of marker genes defined biologically plausible candidates for subsequent studies.

Clustering gene profile vectors has shown that genes encoding proteins with related functions tend to have similar expression patterns (2, 6, 25, 26, 30). For example, one cluster may be associated with genes involved in DNA repair and another with transcription. However, because proteins often have multiple functions and/or roles, it should be possible for a gene profile vector to belong more than one class and for this membership to be quantifiable. The widely used hierarchical clustering approach (6) has sharp rather than smooth cluster boundaries and cannot assign a new profile vector to an existing class. Methods such as SOMs require the number of classes to be specified a priori. The AutoClass (4) implementation of a naive Bayes model clustered the 62 aforementioned experiment profile vectors using a mixture of Gaussian probability distributions and employed Bayesian methods to derive both the maximum posterior probability of classification and the optimum number of classes.

Here, a novel suite of tools developed for the analysis, display, and visualization of genome wide features are applied to 5,687 seventy-eight-feature gene profile vectors from a published *Saccharomyces cerevisiae* molecular profile matrix (25). The eight studies (78 experiments) examined the cell cycle, responses to stress (heat and cold shock), and diauxic shift. AutoClass is

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

*E. J. Moler and D. C. Radisky contributed equally to this work.

Address for reprint requests and correspondence: I. S. Mian, Dept. of Cell and Mol. Biol., MS 74-197, Radiation Biology and Environ. Toxicol. Group Life Sci. Div., Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley CA 94720 (E-mail: SMian@lbl.gov).

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.3 on September 15, 2016

used to discover and characterize classes of gene profile vectors (raw intensity measurements are transformed in a manner that differs from the standard log-ratio). A feature relevance measure different from that proposed earlier (17) is employed to identify experiments that are most important in defining each class. The fraction of genes assigned to a known biological category that belong to a given class is used to pinpoint classes most associated with specific functions. This novel enrichment measure allows external knowledge to be incorporated into the analysis and interpretation procedure in a systematic and quantitative manner. The results are utilized to make inferences about copper and iron homeostasis, a problem that encompasses pathway and mechanisms that were not the primary subject of the original studies. Classes of coexpressed genes suggest those that may be regulated by common factors and thus provide constraints for learning genetic networks within the same graphical model formalism as that used for clustering (17). Future directions for the methodology are discussed.

METHODS

Molecular Profile Matrix

The 78 published cDNA microarray experiments each provide the relative mRNA concentrations for ~6,000 *S. cerevisiae* open reading frames (ORFs) (25). The 8 studies examined the cell division cycle after synchronization with α-factor (abbreviated to Alp, 18 experiments); the cell division cycle measured using a temperature-sensitive *cdc15* mutant (cdc, 25 experiments); the cell division cycle after synchronization by centrifugal elutriation (elu, 14 experiments); sporulation (spo, 10 experiments); diauxic shift (dia, 7 experiments); and three mutants (Clb, 1 experiment; Cln, 2 experiments; and gal, 1 experiment). For 636 ORFs, expression measurements were missing in 7 or more of the 78 experiments, so these were excluded from further analysis. Of the $5{,}687 \times 78 = 443{,}586$ remaining measurements, 2,846 are "missing" data points. MATLAB (**www.mathworks.com**) was employed for all data management, transformation, analysis, visualization and application development. All computations were performed on a Sun Ultra 60 workstation.

A Naive Bayes Model

Graphical models are highly structured stochastic systems that provide a compact, intuitive, and probabilistic framework capable of learning complex relations between variables (for reviews see Refs. (11, 12, and 19) and the introductory tutorial on Bayesian Networks at **www.cs.berkeley.edu/~murphyk/Bayes/bayes.html**). A naive Bayes model is a simple graphical model in which a single unobserved variable C is assumed to generate the observed data (here, 5,687 seventy-eight-feature gene profile vectors). The hidden variable is discrete, and its possible states correspond to the underlying classes in the data. Profile vectors are believed to be generated by K models or data-generating mechanisms. These K models correspond to K clusters or classes of biological interest (Fig. 1). This model-based approach to clustering can handle missing data, noisy data, and uncertainty about class membership in a probabilistic manner. There is direct control over the variability allowed within each class (the variance characteristics of each data-generating mechanism). The question of how many classes the data suggest

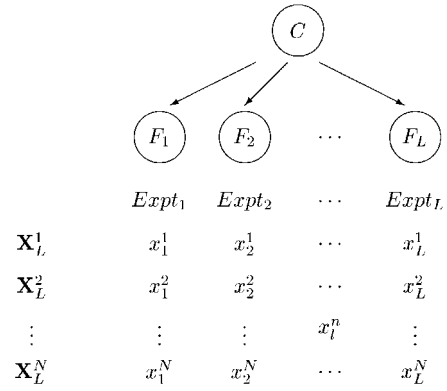

Fig. 1. A graphical model representation of a naive Bayes model and its relationship to N L-feature gene profile vectors. The graph topology is a directed acyclic graph in which nodes represent the variables of interest. C is the hidden classification node that generates K classes present in the $N = 5{,}687$ $L = 78$-feature gene profile vectors ($\mathbf{X}_L^n$ is a gene profile vector). Each F_l node represents the expression levels in experiment l. Influences between variables are encoded explicitly by the presence of edges between nodes. The edges have directionality and thus semantic meaning. The absence of an edge provides information about the independence between concepts: when two variables lack a connecting edge, nothing about the state of one variable can be inferred from the state of the other. The network topology shown makes minimal assumptions about relationships in the data: the experiment F_l variables are conditionally independent given the class C.

can be treated in an objective manner. Finding the optimal weights, locations, and shapes of the component classes can be performed in a principled manner.

The observed data D used to train a model M to discover and characterize classes of gene profile vectors can be represented as $D = [\mathbf{X}_L^1, \ldots, \mathbf{X}_L^N]$, where N and L are the number of genes and experiments respectively. $\mathbf{X}_L^n = [x_1^n, \ldots, x_L^n]$ is an input gene profile vector; x_l^n is the "expression level" of gene n in experiment l calculated in some manner. The section below (*Intensity Transformation*) describes one method for transforming raw intensity measurement to arrive at a value for this quantity. Here, the functional form for the data-generating mechanism is taken to be a Gaussian. For experiment l, the likelihood of expression level x_l^n given class k can be determined from the mean $\mu_{k,l}$ and standard deviation $\sigma_{k,l}$ of the Gaussian modelling class k

$$P(x_l^n|c_{k,l}, M) = \frac{1}{\sqrt{2\pi}\sigma_{k,l}} \exp - \frac{1}{2}\left[\frac{x_l^n - \mu_{k,l}}{\sigma_{k,l}}\right]^2 \qquad (1)$$

Given the relationships depicted in Fig. 1, the likelihood of profile vector $\mathbf{X}_L^n$ given class k is

$$P(\mathbf{X}_L^n|c_k, M) = P(c_k|M) \prod_{l=1}^{L} P(x_l^n|c_{k,l}, M) \qquad (2)$$

where $P(c_k|M)$ is the prior probability of class k, $\sum_{k=1}^{K} P(c_k|M) = 1$ and $0 \leq P(c_k|M) \leq 1$. The likelihood of the profile vector given the model $P(\mathbf{X}_L^n|M)$ is a sum over all K classes. If the N genes are assumed to be identical and independently distributed, then the likelihood of the data D given model M is

$$P(D|M) = \prod_{n=1}^{N} P(\mathbf{X}_L^n|M) = \prod_{n=1}^{N} \sum_{k=1}^{K} P(\mathbf{X}_L^n|c_k, M) \qquad (3)$$

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.3 on September 15, 2016

Given only the observed data D and a functional form for the mechanism that generates D, the task is to find a model M that best describes D and hence the classes. Since the network topology shown in Fig. 1 is fixed, the learning problem becomes one of estimating the parameters of M. For L observed variables, these are the number of classes K and $K \times L$ local probability models and conditional probability distributions. Here, the probability parameters are the Gaussian mean and standard deviation, so $K \times L \times 2$ parameters need to be determined. Estimating these parameters from data involves finding a maximum a posteriori model: a model which maximizes the posterior probability of the model given the data (Bayes rule)

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)} \tag{4}$$

If the space of models is assumed to be fixed, then $P(D)$ is constant. Introducing a uniform prior over models, assuming that all models are equally likely, leads to $P(M|D) \propto P(D|M)$. Thus, model estimation reduces to finding the model that maximizes $P(D|M)$ (*Eq. 3*). A trained model assigns a profile vector to one of K existing classes by determining the class which maximizes *Eq. 2*.

AutoClass Implementation of Naive Bayes Models

AutoClass C version 3.3 (4) models the continuous experiment F_l nodes in Fig. 1 using Gaussians and the discrete classification C node using a Bernoulli distribution. Starting from random initial descriptions for a specified number of classes, a gradient descent search through the space of descriptors is performed. At each step of the search procedure, the current descriptions are used to assign probabilistically each profile vector to each class. The observed values for each profile vector are used to update class descriptions, and the procedure is repeated until a specified convergence criterion is reached. A variant of the expectation-maximization (EM) algorithm is employed with the additional assumption that each profile vector belongs to some class (the sum of all class probabilities is one). There is a penalty for adding more classes and thus overfitting the data. Increasing the number of classes will decrease the prior probability of each class unless the additional class improves the likelihood of the data (*Eq. 3*). AutoClass iterates through different numbers of classes to determine the best taxonomy.

The model-space that needs to be searched can be constrained by setting a lower bound on the variance of the data-generating mechanism. For each experiment l, the level of observation noise (measurement error) and/or the natural variation in expression between genes can be used to set this value in a data-dependent manner. However, since neither the noise nor intrinsic variability are known, a lower bound on the standard deviations of Gaussians modeling classes is set to 1/10 of the standard deviation of N expression levels $\{x_l^1, \ldots, x_l^N\}$. Thousands of models are estimated, each starting from different random number seeds. Each resultant model, a locally optimum solution in the parameter space, is scored. The model marginals are compared to find the model that best describes the data. The results do not depend on the order in which profile vectors are entered into a model.

The input data for AutoClass are the 5,687 seventy-eight-feature gene profile vectors $[\mathbf{X}_{78}^1, \ldots, \mathbf{X}_{78}^{5687}]$, where $\mathbf{X}_{78}^n = [x_1^n, \ldots, x_{78}^n]$. The output consists of *1*) K, the number of classes, *2*) an $N \times K$ likelihood matrix where each element is the likelihood of a gene profile vector n given class k $P(\mathbf{X}_L^n|c_k, M)$, and *3*) a $K \times L$ parameter matrix where each element is the mean and standard deviation of the Gaussian modeling class k and experiment l, $(\mu_{k,l}, \sigma_{k,l})$. The marginal for the best model, the one used in all subsequent analyses, is significantly higher than the next nine models. Since all $L = 78$ experiments are used, the gene profile vector classes discovered and characterized by the best model capture the expression behavior of genes across this specific range of conditions. Although some of the 8 studies consisted of more than one experiment, the topology of the naive Bayes model treats the 78 experiments as being independent. The notion of conditional independence of gene profile vectors given a class does not mean independence from the experimental conditions, an unlikely assumption (1).

Intensity Transformation

Frequently, the "expression level" of gene n in experiment l, x_l^n, is taken to be the log of the background-corrected intensity measurements for samples tagged with the Cy5 and Cy3 dyes, $x_l^n = \log(I_{\mathrm{Cy5}}/I_{\mathrm{Cy3}})$. Here, the expression level is calculated by normalising background corrected intensities using $x_l^n = (I_{\mathrm{Cy5}} - I_{\mathrm{Cy3}})/(I_{\mathrm{Cy5}} + I_{\mathrm{Cy3}})$. This transformation has the same advantages as the log-ratio but has two additional useful properties. It minimizes errors associated with background subtraction from low-intensity signals and constrains expression levels to lie in the $-1 \leq I \leq +1$ domain. The expression levels in the gene profile vectors used to train a naive Bayes model are not shifted, rescaled, or modified in any other way.

AutoClass Implementation of Feature Relevance

AutoClass implements a feature relevance measure termed the relative influence I_l^k. Here, it signifies how important experiment l is in determining class k. It is defined as the relative entropy between two distributions representing the expression level for the N genes in the trained model $P(x_l^n|c_{k,l}, M)$ and a model M^* describing the entire data set by means of a single class $P(x_l^n|M^*)$

$$I_l^k \equiv \sum_{n=1}^{N} P(x_l^n|c_{k,l}, M) \log \frac{P(x_l^n|c_{k,l}, M)}{P(x_l^n|M^*)} \tag{5}$$

For experiment l, the Gaussian mean and standard deviation for a single class can be calculated directly from $\{x_l^1, \ldots, x_l^N\}$. This allows terms involving M^* to be computed using *Eq. 1*. Reordering the L experiments according to their relative influence values ranks them in terms of their importance in defining class k. The most (least) influential experiment is one that maximizes (minimizes) I_l^k.

Integrating External Knowledge to Aid Interpretation: Enrichment

To create a systematic and quantitative environment in which to interpret naive Bayes model classes, external knowledge about genes is integrated in the following manner. An $\Omega \times K$ enrichment matrix is calculated as the product of an $\Omega \times N$ ontology matrix and the $N \times K$ AutoClass likelihood matrix (see above, *AutoClass Implementation of Naive Bayes Models*). Given a specific type of external knowledge, let α be one of the Ω categories to which the ontology assigns a gene. For example, an ontology could classify genes based upon cell type, developmental stage, and so on. Each element of an ontology matrix is the probability of gene n given category α, $P(X^n|\alpha)$. The enrichment matrix is a product of the ontology and likelihood matrices normalized so that the sum across all classes in a single category is one. Each

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.3 on September 15, 2016

element of this enrichment matrix, ϵ_α^k or enrichment, is the fraction of genes in category α that belong to class k

$$\epsilon_\alpha^k \equiv \frac{\sum_{n=1}^{N} P(\mathbf{X}_L^n|c_k, M)P(X^n|\alpha)}{\sum_{k=1}^{K}\sum_{n=1}^{N} P(\mathbf{X}_L^n|c_k, M)P(X^n|\alpha)} \quad (6)$$

where $\sum_{k=1}^{K} \epsilon_\alpha^k = 1$ and $0 \leq \epsilon_\alpha^k \leq 1$. Reordering classes according to their enrichment values allows them to be ranked in terms of their association with a category. The most (least) enriched class is the one which maximizes (minimizes) ϵ_α^k.

Four enrichment matrices are calculated using ontologies that assign genes to the following sets of categories: *1*) protein function, $\Omega = 106$ highest level categories in the MIPS functional catalog (16); *2*) M/G_1, G_1, S, S/G_2, and G_2/M stage of the cell cycle, $\Omega = 5$ stages (this matrix uses only the 800 genes identified using Fourier analysis of the cell cycle experiments (25) as being most associated with these stages); *3*) target of transcription factor, $\Omega = 265$ categories given by the YPD version 9.36a (10); and *4*) chromosome number, $\Omega = 16$ categories based upon chromosomal location given by the YPD version 9.36a (10). For simplicity, the assignment of a gene to a category is taken to binary: $P(X^n|\alpha)$ is set to 1.0 (0.0) if the gene is (is not) a member of category α.

RESULTS

Gene Profile Vector Classes

A naive Bayes model trained using 5,687 seventy-eight-feature gene profile vectors identified 45 classes (referred to as *classes 1–45*). Four exemplars of these classes are shown in Fig. 2. Genes for which the class probability (likelihood) is less than 1.0 have some probability of belonging to at least one other class (see *class 1*). The boundary for a class can be sharp if all its members have a probability of 1.0 (see *class 40*). *Class 40* shows an example of a class pattern that captures the periodicity of time series studies. In *class 42,* the

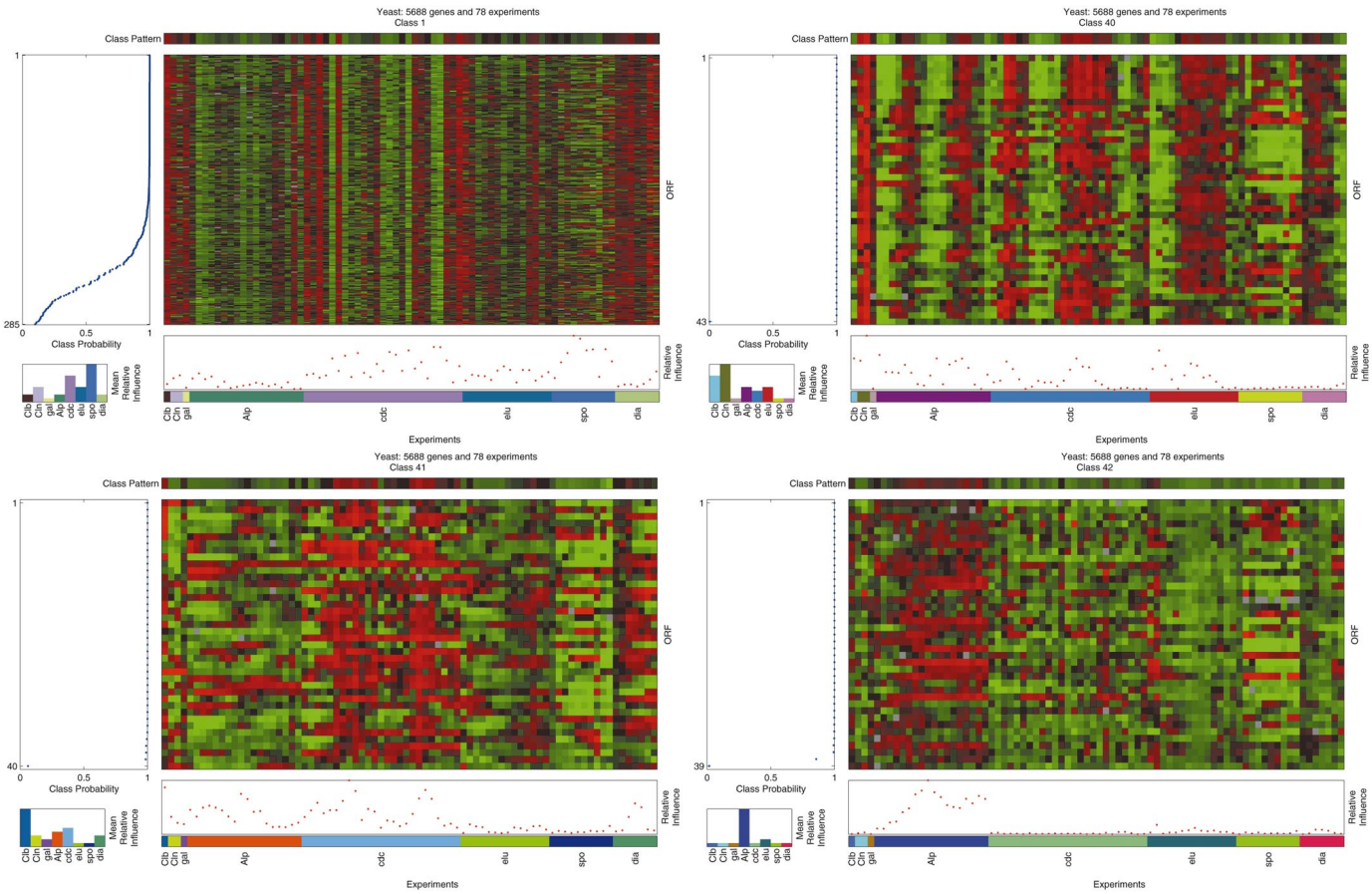

Fig. 2. The general format devised to display the K classes discovered and characterized by a naive Bayes model trained using N L-feature gene profile vectors. Shown are four of the $K = 45$ classes estimated from 5,687 seventy-eight-feature yeast gene profile vectors: *class 1* (*top left*), *class 40* (*top right*), *class 41* (*bottom left*), and *class 42* (*bottom right*). Each column represents one of the $L = 78$ experiments, and every row represents one of the $N = 5{,}687$ genes. Each element represents the expression level of a gene color-coded on a scale where red (green) signifies increased (decreased) expression. The order of experiments is indicated by the colored bar (the Alp, elu, cdc, spo, and dia studies are all time series). The genes are sorted according to their "class probability" or likelihood of the gene profile vector given class k and model M, $0.0 \leq P(\mathbf{X}_L^n|c_k, M) \leq 1.0$ (see *A Naive Bayes Model, Eq. 2*). For brevity, only the 285, 43, 40, and 39 genes with the highest class probabilities given *classes 1, 40, 41,* and *42,* respectively, are shown. "Class pattern" represents the means of the Gaussians modeling the 78 experiments in class k (a "prototypical" gene profile vector). "Relative influence," I_l^k, signifies how influential experiment l is in defining class k (see *AutoClass Implementation of Naive Bayes Models, Eq. 5*). "Mean relative influence" is the average of the relative influence for studies that contain more than one experiment.

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.3 on September 15, 2016

relative influence and mean relative influence values indicate that the study most important in defining this class is the Alp time series (this class is discussed in more detail subsequently). The number of genes assigned to each class can be approximated by the Class weight, $w_k = \Sigma_{n=1}^{N} P(\mathbf{X}_L^n|c_k, M)$. The classes and their weights are as follows 1:244, 2:228, 3:222, 4:218, 5:210, 6:189, 7:184, 8:173, 9:166, 10:161, 11:158, 12:156, 13:154, 14:152, 15:147, 16:145, 17:145, 18:144, 19:142, 20:142, 21:139, 22:135, 23:131, 24:128, 25:118, 26:115, 27:115, 28:115, 29:109; 30:107, 31:100, 32:98, 33:98, 34:91, 35:88, 36:84, 37:79, 38:76, 39:57, 40:42, 41:39, 42:38, 43:36, 44:36, and 45:33.

Of the 33 genes assigned to *class 45* (data not shown), 5 are associated with cell cycle control (PCL5), energy generation (HAP1, SHY1), and amino acid metabolism (ARG1, CPA2). Their YPD annotations are ARG1 [YOL058W; argininosuccinate synthetase (citrulline-aspartate ligase); catalyses the penultimate step in arginine synthesis]; CPA2 [YJR109C; carbamoylphosphate synthase of arginine biosynthetic pathway, synthetase (large) subunit]; HAP1 (YLR256W; transcription factor with heme-dependent DNA-binding activity; responsible for heme-dependent activation of many genes); PCL5 (YHR071W; cyclin that associates with Pho85p); and SHY1 (YGR112W; mitochondrial protein required for respiration). It remains to be seen whether the 28 members with an "unknown" YPD role have one or other of these biological functions ("guilt-by-association").

Association Between Classes and Gene Categories: Enrichment

Figure 3 shows the relationship(s) between gene profile vector classes and four types of external knowledge. Because of limitations in how enrichment values are computed, subsequent discussions will focus on a qualitative assessment of selected observations that highlight the overall utility of enrichment matrices. For example, identical enrichment values can be obtained that do not have the same significance, and there is no correction for classes having different numbers of members. In addition, the assignment of genes to categories is neither comprehensive nor complete, so it is difficult to estimate the actual false positive and false negative rates. The emphasis will be on classes shown in Fig. 2.

Protein function. Class 40 (42 members) is most associated with 32 genes assigned to "CELLULAR ORGANIZATION; organization of chromosome structure". Similarily, *Class 41* (39 members) is associated with 11 genes assigned to "METABOLISM; phosphate metabolism; phosphate utilization". If a single class is associated primarily with a single category, then it is possible to suggest a biological function for genes assigned to the class that have an "unknown" YPD role. For example, *Class 25* contains a large proportion of genes in the category "PROTEIN SYNTHESIS; ribosomal proteins". This class contains "translationally controlled tumor protein" (TCTP), a highly conserved eucaryotic cytoplasmic protein found in several normal and tumor cells that is believed to have a general, yet unknown, housekeeping function (22). Since the likelihood of yeast TCTP (YKL056C) given *class 25* is 1.0, this protein may have a ribosomal function. A general biological function subdivided into a set of categories can be segregated amongst many classes. For example, "ENERGY" is associated with *class 43* (ENERGY; fermentation, ENERGY; gluconeogenesis, ENERGY; glycolysis), *class 31* (ENERGY; glyoxylate cycle) and Class 30 (ENERGY; tricarboxylic acid pathway).

M/G_1, G_1, S, S/G_2, and G_2/M stage of the cell cycle. These 5 categories are associated with only a subset of the 45 classes. For example, M/G_1 is associated with *class 39,* but genes assigned to G_1 are distributed across *classes 26* and *37.* The results indicate that one class (category) can be associated with more than one category (classes).

Target of transcription factor. Classes 29, 39, 40, 41, 42, 43, and *44* are associated with specific transcription factors. For *class 40,* these factors regulate cell cyclins and chromatin assembly genes (BCK2, HTA1, HTA2, HTB1, SIT4, SPT5, SPT6, SPT10, SPT12). This observation is consistent with the protein function category most associated with *class 40* being "CELLULAR ORGANIZATION; organization of chromosome structure". Thus, members of *class 40* with "unknown" YPD roles may be involved directly or indirectly in cellular organization and could be regulated by the aforementioned transcriptions factors. These ORFs are YDR451C, YKR012C, YMR215W, YMR305C, YNL300W, YNR009W, YOL007C, YOL019W, YOR247W, and YOR248W. A similar type of prediction can be made for ORFs in *class 43:* YCR013C and YKL153W may be involved in energy metabolism and could be regulated by PDC2 (pyruvate decarboxylase regulatory protein), GCR2 (transcriptional activator involved in regulation of glycolytic gene expression), and REG1 (regulatory subunit for protein phosphatase Glc7p required for glucose repression).

Chromosome number. At the level of resolution of entire chromosomes, there is little association between class and chromosome number. It is conceivable that partitioning the genome into smaller segments might reveal classes associated with specific regions of a chromosome and thus potential common noncoding regulatory regions.

Genes Assigned to More than One Class

Table 1 of the Supplementary Material lists genes that best illustrate partial membership of multiple classes. (Tables 1–4 have been published online as Supplementary Material and can be viewed at the *Physiological Genomics* web site.)[1] The genes that are almost equally distributed between the largest

[1]Supplementary Material to this article (Tables 1–4) is available online at **http://physiolgenomics.physiology.org/cgi/content/full/4/2/127/DC1**.

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.3 on September 15, 2016

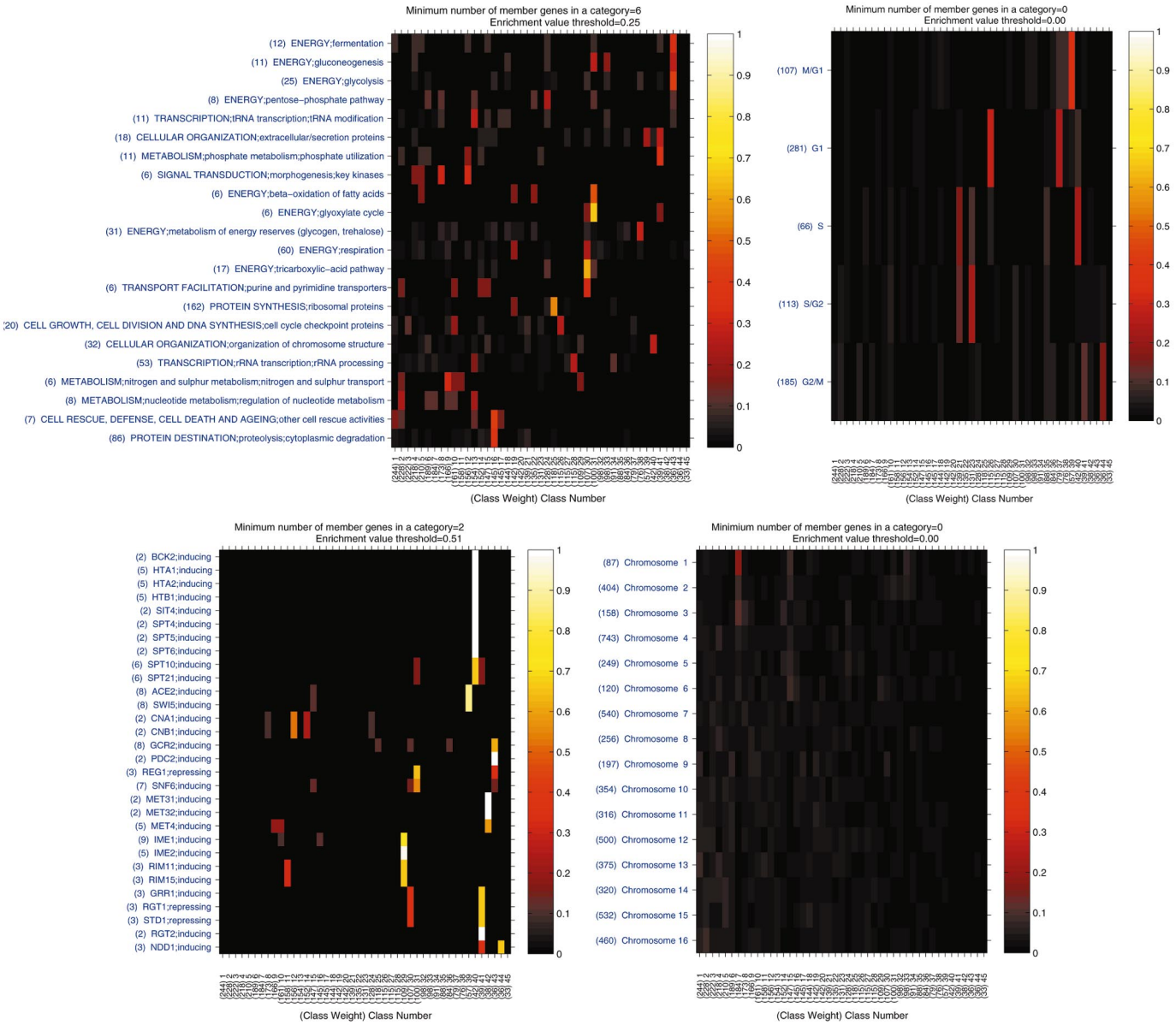

Fig. 3. The general format devised to display the relationships between the profile vector classes (columns) and external information (rows). The four enrichment matrices shown are "protein function" (*top left*), M/G_1, G_1, S, S/G_2, and G_2/M stage of the cell cycle (*top right*), "target of transcription factor" (*bottom left*) and "chromosome number" (*bottom right*) (see *Integrating External Knowledge to Aid Interpretation: Enrichment*). Each element represents enrichment, the degree of association between genes assigned to category α and class k. The values, $0.0 \leq \epsilon_\alpha^k \leq 1.0$, are color coded according the scale shown by the vertical bar (yellow-white signifies the greatest association, highest value). For example, the 12 genes in the protein function category "ENERGY:fermentation" are most associated with *class 43* (*top right*). The rows of the enrichment matrices have been filtered to display only those categories that have more than a minimum number of member genes and/or enrichment values above a threshold. "Class weight" approximates the number of genes assigned to each of the $K = 45$ naive Bayes model classes.

number of classes are YLR020C, YLR113W (HOG1), and YOR023C (AHC1). For YLR160C (ASP3D) and YHR004C (NEM1), not only is the maximum value of the likelihood of the profile vector given a class less than 0.5, but they are assigned to 5 different classes (the maximum possible). For genes that belong to multiple classes, the eight YPD cellular roles assigned to two or more genes are *Pol* II transcription, protein modification, cell stress, small molecule transport, vesicular transport, signal transduction, amino acid metabolism, and protein degradation.

For genes in Table 1 (of the Supplementary Material) with known YPD roles, the annotations for other genes assigned to the same set of classes was examined. This analysis yields some novel and interesting observa-

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.3 on September 15, 2016

tions, especially with respect to yeast genes important in regulating metal transport (reviewed in Ref. (20)). YAL021C (CCR4) is assigned to *classes 2, 5,* and *7:* these classes contain many other genes that have been linked genetically with CCR4. YGL233W (SEC15) and YFL025C (BST1) are assigned to *classes 2* and *7:* these classes contain other genes shown to be involved in vesicle transport (SEC15 also falls into *class 4*). YGL167C (PMR1), which encodes a protein involved in import of calcium, copper, and manganese into the Golgi, is assigned to *classes 8, 12, 14,* and *24:* these classes contain many other genes known to be involved in intracellular metal metabolism. For example, *classes 8* and *24* contain GEF1, a gene encoding an intracellular chloride transporter essential for proper assembly of copper into the iron transport protein Fet3p. *Class 12* contains MMT2, which encodes a mitochondrial iron transporter that results in enhanced survival on low iron conditions. *Class 14* contains both YGL071W (RCS1/AFT1), which encodes the key transcriptional regulator of iron metabolism, and BSD2, which encodes a posttranslational regulator of Smf1p, a manganese transporter. PBN1, RPN4, ASP3, HOG1, and CIN1 are genes with complex functions that are likely to be associated with many pathways and/or interactions.

To predict potential biological functions for genes in Table 1 with an "unknown" YPD role, other genes that exhibit the same spectrum of class memberships were identified (see Table 2 of the Supplementary Material). Comparing these two tables suggests that these eight genes could be involved in *Pol* II transcription, protein degradation, cell stress, amino acid metabolism, lipid, fatty acid, and sterol metabolism, and RNA turnover.

Genes Involved in Copper and Iron Metabolism

The original studies were designed to reveal expression patterns associated with the cell cycle, responses to stress, and diauxic shift. It is not surprising, therefore, that many of the classes identified here are highly enriched in genes involved in processes such as release or storage of energy (see Fig. 3); these pathways are likely to be coordinated with periods of the cell cycle in which varying amounts of energy is required. An important issue is whether analysis of this same data set can reveal other, potentially unknown regulatory pathways that might be less dependent upon the cell cycle. To address this question, genes involved in metal transport were selected for further study. Many yeast genes involved in metal metabolism have been characterized, and a priori such genes might not be expected to have a strong cell cycle dependency. Although yeast has proven to be a valuable model system for identification and characterization of biological processes relevant to a number of human diseases related to metal metabolism, many aspects of metal physiology remain to be discovered.

Genes known to be important in iron and copper metabolism were identified and examined in more detail (see Table 3 in the Supplementary Material). Classes which contain more than one of these genes are as follows: *class 4,* TAF17 ($P(\mathbf{X}_L^n|M)$ = 0.88); FRE4 (0.47); *class 5,* GEF1 (0.02), MMT1 (0.02); *class 6,* LYS7 (1.00), YFH1 (0.03); *class 8,* GEF1 (0.91), MNR2 (0.01); *class 9,* SMF2 (1.00), CTR2 (1.00), YAH1 (1.00), TAF19 (0.07), TAF17 (0.01); *class 12,* MAC1 (1.00), MMT2 (1.00); *class 14,* RCS1 (0.99), TAF145 (1.00); *class 18,* SMF1 (1.00), FRE7 (1.00); *class 20,* ATX1 (1.00), ISU2 (1.00); *class 22,* SLF1 (1.00), CRT2 (1.00), CRS5 (1.00), MNR2 (0.94); *class 30,* RIP1 (1.00), CUP5 (1.00), FET5 (1.00), SDH2 (1.00), ISU1 (0.99), FTH1 (1.00); *class 34,* ATM1 (1.00), GEF1 (0.02); *class 35,* CUP1A (1.00), CUP1B (1.00); *class 41,* FTR1 (1.00), FET3 (1.00); and *class 42,* SIT1 (1.00), ARN1 (1.00), TAF1 (1.00), FRE6 (1.00), FRE1 (1.00), ENB1 (1.00), CTR1 (1.00). There are numerous instances in which genes with similar functions are assigned to the same class. For example, SMF2 and CTR2, genes which encode low-affinity transporters of manganese and copper, respectively, are assigned to *class 9.* FET3 and FTR1, which together encode the yeast high-affinity iron transporter, is assigned to *class 41.*

The 38 members of *class 42* are of particular interest since 7 are known to be involved in copper and iron transport (20). The proteins encoded by CTR1, FRE1, and FRE6 have all been implicated in copper uptake. SIT1, which encodes a protein involved in uptake of siderophore-bound iron, is classified with three closely related genes (ARN1, TAF1, and ENB1) that are known to be tightly regulated by iron need and that are collectively required for normal growth on low-iron medium (31). It remains to be seen whether the three transcription factors most associated with *class 42* (Fig. 3) do indeed play a role in regulating copper and iron transport. These transcription factors are Met31 and Met32 (zinc-finger proteins involved in transcriptional regulation of methionine metabolism) and Met4 (transcriptional activator of the sulfur assimilation pathway).

Table 4 of the Supplementary Material shows the remaining 31 members of *class 42,* some of which have known roles. YDR040C (ENA1) is required for high-salt tolerance. YDR340W is similar to HAP1, a gene which encodes a complex transcriptional regulator of many genes involved in electron-transfer reactions and which is essential in anaerobic or heme-depleted conditions. These data suggest that the 17 members with an "unknown" YPD role represent good candidates for new genes involved in metal metabolism, especially copper and iron homeostasis. The precise mechanisms by which they regulate and maintain metal ion homeostasis and the pathways in which they participate can only be inferred by the function of the other members of *class 42.* The regulatory mechanism(s) that unites members of *class 42* is as yet undiscovered, although one good possibility is that it may involve the activity of the transcription factors encoded by MET4, MET31, and MET32. Specific experiments using yeast mutants with deletions of genes listed in Table 4 could help clarify the situation as well as suggest additional new candidates. Interestingly, two of the uncharac-

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.3 on September 15, 2016

terized ORFs have human homologs (YJR033C, YDR534C), suggesting that as yet unidentified human disorders may result from aberrant regulation or functioning of these proteins.

The ability of the current analysis to provide insights into unknown regulatory mechanisms relevant to metal metabolism suggests that other classes which contain a high proportion of unknown ORFs could be used to investigate other physiological processes.

DISCUSSION

The cornucopia of transcription profile data and other information available for *S. cerevisiae* makes it an excellent model system for investigating cellular metabolic processes. This work addressed the problem of extracting statistically and biologically meaningful insights from such data in a systematic manner. A naive Bayes model was used to discover and characterize classes of gene profile vectors. The probability parameters of the 45 classes were combined with external knowledge to determine the relationship between each class and a particular biological category. By suggesting, for example, specific transcription factors that are most associated with each class, new experiments can be designed aimed at identifying common regulatory mechanisms. The techniques described here provide a method for predicting potential functions of currently uncharacterized genes and their products based on similarity of global gene expression patterns with known roles rather than similarity to structures and/or sequences. Thus, although the translationally controlled tumor protein exhibits no obvious sequence similarity to known ribosomal or ribosome-associated proteins, its pattern of gene expression across the experiments examined suggests that it might be associated with this physiological role.

The published studies characterized genes involved in several housekeeping functions. Currently available hierarchical clustering methods identified a large number of genes associated with the cell cycle but did not focus on associations between genes involved in other biological pathways and metabolic functions. The results here suggest a number of connections to genes involved in copper and iron homeostasis even though a link between metal metabolism and the cell cycle might not be expected a priori. The 17 ORFs in *class 42* predicted to be involved directly or indirectly in these processes are good targets for investigation by gene deletion and/or other techniques. Since many genes involved in metal metabolism in yeast have human homolog that are altered in disease states, studies of these new yeast candidates may yield unanticipated information on biochemical mechanisms relevant to human disorders. Using techniques different from those employed here, Tavazoie et al. (27) addressed the issue of identifying other pathways and noncoding regulatory motifs (an area not considered in this work). The association between *class 42* and the MET4, MET31, and MET32 transcription factors suggests that this class may be equivalent to *cluster 30* in their work. A more complete comparison of their MIPS functional category enrichment with the enrichment measure computed here may be warranted.

In the graphical model used for clustering gene profile vectors, experiments were treated as independent of each other. Despite the simple nature of this probabilistic model and the assumption of a Gaussian functional form for the data-generating mechanism, novel, biologically plausible observations could be made. Consequently, future experiments of particular interest include comparisons of wild-type yeast with single and multiple deletion mutants of genes predicted here to be involved in copper and iron homeostasis as well as those known to play roles in these processes. The utility of gene profile vector classes could be enhanced by computing a hierarchical structure capturing the relationships between classes, i.e., clustering the classes. As the number and variety of experiments performed increases, a straightforward extension is two-way clustering, finding classes of experiment and gene profile vectors together with associations between them. For the same data set, it will be useful to compare the posterior probabilities of a K-class naive Bayes model with a K-class (Gaussian) mixture model computed from the clusters generated by a K-class SOM.

For the time series studies, the independence supposition may be erroneous, since the measured expression level at one time point (modeled by node F_l) might be related to those in one or more previous time points ($F_{j<l}$). Such temporal dependencies could be encoded in the topology of the graphical model by adding edges between time points (Fig. 4). Nodes could be included that represent gene profile vector labels such as biochemical activity, protein fold, and so on. An important area of future research is extending the graphical model formalism used for clustering to graphical models for inferring genetic networks. For example, the "cdc" time series experiments could be modeled using dynamic Bayesian networks (18).

Here, the functional form for the data-generating mechanism was a Gaussian. Even if the observed distribution of gene expression levels is unimodal, it may not be Gaussian. As the number of experiments increases, the data themselves could be used to estimate more appropriate nonparametric and/or semi-parametric functional forms. In modeling amino acid distribu-

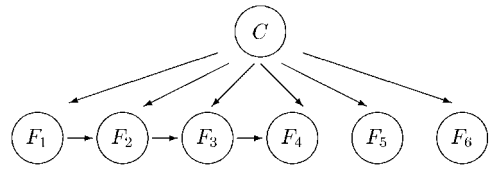

Fig. 4. Clustering gene profile vectors using a graphical model that takes into consideration correlations between variables. In the example shown, the first four experiments represent a time series study. The joint distribution is $P(F_1, \ldots, F_6, C|M) = P(C|M)P(F_1|C, M)\,\Pi_{l=2}^{l=4} P(F_l|F_{l-1}, C, M)\,\Pi_{l=5}^{l=6} P(F_l|C, M)$. If the F_l nodes are independent given the node C, then the joint distribution becomes $P(F_1, \ldots, F_6|M) = P(C|M)\,\Pi_{l=1}^{l=6} P(F_l|C, M)$.

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.3 on September 15, 2016

tions in protein multiple sequence alignments, for example, a mixture of Dirichlet densities estimated from large databases of multiple sequence alignments were combined with observed amino acid frequencies to form estimates of expected amino acid probabilities at each position in a profile, hidden Markov model, or other statistical model (23). Eventually, it should be possible to develop analogous mixture models and priors in which a given observed profile vector is assumed to be a mixture of prototypical profile vectors estimated from a diverse array of experiments. These prototypical distributions could be regarded as different states of the system under study.

This work was supported by a Hollaender distinguished postdoctoral fellowship (to D. C. Radisky) and the Director, Office of Science, Office of Biological and Environmental Research, Life Sciences Division (to I. S. Mian and E. J. Moler) under US Department of Energy Contract No. DE-AC03-76SF00098. The data are available upon request.

Present address of E. J. Moler: Chiron Corp., 4560 Horton St., Emeryville CA 94608.

REFERENCES

1. **Aach J, Rindone W, and Church GM.** Systematic management and analysis of yeast gene expression data. *Genome Res* 10: 431–445, 2000.
2. **Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, and Levine AJ.** Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA* 96: 6745–6750, 1999.
3. **Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Jr, and Haussler D.** Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA* 97: 262–267, 2000.
4. **Cheeseman P and Stutz J.** Bayesian classification (AutoClass): theory and results. In: *Advances in Knowledge Discovery and Data Mining,* edited by Fayyad UM, Piatetsky-Shapiro G, Smyth P, and Uthurusamy R. AAAI Press/MIT Press, 1996. [The software is available at http://ic-www.arc.nasa.gov/ic/projects/bayes-group/autoclass/index.html]
5. **D'Haeseleer P, Wen X, Fuhrman S, and Somogyi R.** Linear modeling of mRNA expression levels during CNS development and injury. In: *Pacific Symposium on Biocomputing,* 1999, p. 41–52.
6. **Eisen MB, Spellman PT, Brown PO, and Botstein D.** Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95: 14863–14868, 1998.
7. **Friedman N, Linial M, Nachman I, and Pe'er D.** Using Bayesian networks to analyze expression data [Online]. Stanford University. http://robotics.stanford.edu/people/nir/publications.html [2000].
8. **Goss PJ and Peccoud J.** Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets. *Proc Natl Acad Sci USA* 95: 6750–6755, 1998.
9. **Hastie T, Tibshirani R, Eisen M, Brown P, Ross D, Scherf U, Weinstein J, Alizadeh A, Staudt L, and Botstein D.** Gene shaving: a new class of clustering methods for expression arrays [Online]. Stanford University. http://www-stat.stanford.edu/~hastie/Papers/ [2000].
10. **Hodges PE, McKee AH, Davis BP, Payne WE, and Garrels JI.** The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res.* 27: 69–73, 1999. [The database is available at http://www.proteome.com/databases/index.html]
11. **Jensen VF.** An introduction to Bayesian Networks. London: UCL, 1996.
12. **Jordan MI** (editor). *Learning in Graphical Models.* Dordrecht, Netherlands: Kluwer Academic, 1998.
13. **Kauffman S.** *The Origins of Order. Self-Organization and Selection in Evolution.* Oxford: Oxford University Press, 1993.
14. **Liang S, Fuhrman S, and Somogyi R.** Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In: *Pacific Symposium on Biocomputing,* 1998, p. 18–29.
15. **Matsuno H, Doi A, Nagasaki M, and Miyano S.** Hybrid Petri net representation of gene regulatory network. In: *Pacific Symposium on Biocomputing*, 2000, vol. 5, p. 338–349.
16. **Mewes HW, Heumann K, Kaps A, Mayer K, Pfeiffer F, Stocker S, and Frishman D.** MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 27: 44–48, 1999.
17. **Moler EJ, Chow ML, and Mian IS.** Analysis of molecular profile data using generative and discriminative methods. *Physiol Genomics* 4: 109–126, 2000.
18. **Murphy K and Mian IS.** Modelling gene expression data using dynamic Bayesian networks [Online]. University of California, Berkeley. http://www.cs.berkeley.edu/~murphyk/publ.html [1999].
19. **Pearl J.** *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1988.
20. **Radisky D and Kaplan J.** Regulation of transition metal transport across the yeast plasma membrane. *J Biol Chem* 274: 4481–3384, 1999.
21. **Raychaudhuri R, Stuart JM, and Altman RB.** Principal components analysis to summarize microarray experiments: application to sporulation time series. In *Pacific Symposium on Biocomputing*, 2000, vol. 5, p. 452–463.
22. **Sanchez JC, Schaller D, Ravier F, Golaz O, Jaccoud S, Belet M, Wilkins MR, James R, Deshusses J, and Hochstrasser D.** Translationally controlled tumor protein: a protein identified in several nontumoral cells including erythrocytes. *Electrophoresis* 18: 150–155, 1997.
23. **Sjölander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, and Haussler D.** Dirichlet mixtures: a method for improving detection of weak but significant protein sequence homology. *CABIOS* 12: 327–345, 1996. [More up-to-date information is available at http://www.cse.ucsc.edu/research/compbio/dirichlets/index.html]
24. **Somogyi R and Sniegoski CA.** Modeling the complexity of genetic networks: understanding multigenetic and pleiotrophic regulation. *Complexity* 1: 45–63, 1996.
25. **Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, and Futcher B.** Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 9: 3273–3297, 1998. [The data are available at http://cellcycle-www.stanford.edu]
26. **Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, and Golub TR.** Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* 96: 2907–2912, 1999.
27. **Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, and Church GM.** Systematic determination of genetic network architecture. *Nat Gen* 22: 281–285, 1999.
28. **Toronen P, Kolehmainen M, Wong G, and Castren E.** Analysis of gene expression data using self-organizing maps. *FEBS Lett* 451: 142–146, 1999.
29. **Weaver DC, Workman CT, and Stormo GD.** Modeling regulatory networks with weight matrices. In: *Pacific Symposium on Biocomputing,* 1999, p. 112–123.
30. **Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, and Somogyi R.** Large-scale temporal gene expression mapping of central nervous system development. *Proc Natl Acad Sci USA* 95: 334–339, 1998.
31. **Yun C-Y, Ferea T, Rashford J, Ardon O, Brown PO, Botstein D, Kaplan J, and Philpott CC.** Desferioxamine-mediated iron uptake in *Saccharomyces cerevisiae:* evidence for two pathways of iron uptake. *J Biol Chem* In press.

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.3 on September 15, 2016