BIOINFORMATICS

Vol. 1 no. 1 2003
Pages 1–9

Discovering Molecular Pathways from Protein Interaction and Gene Expression data

E. Segal[1], H. Wang[1] and D. Koller[1]

[1]Computer Science Dept., Stanford University, Stanford, CA 94305-9010, USA

ABSTRACT
In this paper, we describe an approach for identifying "pathways" from gene expression and protein interaction data. Our approach is based on the assumption that many pathways exhibit two properties: their genes exhibit a similar gene expression profile, and the protein products of the genes often interact. Our approach is based on a unified probabilistic model, which is learned from the data using the EM algorithm. We present results on two *S. Cerevisiae* gene expression data sets, combined with a binary protein interaction data set. Our results show that our approach is much more successful than other approaches at discovering both coherent functional groups and entire protein complexes.
Contact: eran@cs.stanford.edu
Keywords: probabilistic models, protein interaction, gene expression.

INTRODUCTION

Cellular processes are carried out through interactions among many genes and gene products. This activity is often organized into *pathways*: sets of genes that coordinate to achieve a specific task. Revealing this organization is crucial to obtaining a coherent global picture of cellular activity.

Recent technological advances enable us to extract many different types of genomic data, including: DNA sequences, gene expression measurements, protein-protein interactions, and DNA binding data. These data provide us for the first time with the means to get at the modular organization of the cell on a genome wide scale. Indeed, much recent work has been devoted to the analysis of these data for this purpose. However, most of this work has been devoted to the analysis of a single type of data, using other types of data only for validation. In this paper, we propose an integrated approach that attempts to discover pathways using both gene expression and protein-protein interaction data.

Our approach is based on the assumption that many pathways exhibit two properties. First, genes in the same pathway are activated together, and thus exhibit similar gene expression profiles. Second, when genes coordinate to achieve a particular task, their protein products often interact. The asssumption that many pathways exhibit both properties is supported by the work of Ge *et al.* (2001), that showed that genes with similar expression profiles are more likely to encode interacting proteins, and by the study of Jansen *et al.* (2002), which showed that the genes of experimentally derived protein complexes are often co-expressed.

Based on this assumption, our approach searches for sets of genes that have a similar expression profile, and a significant fraction of protein-protein interactions in the DIP binary interaction data set (Xenarios *et al.*, 2000). This unified approach has advantages over approaches that use only a single type of data. For example, many analyses use clustering to construct groups of genes that have similar expression profiles (e.g., (Alon *et al.*, 1999; Eisen *et al.*, 1998)). However, it is often hard to conclude that the resulting clusters actually correspond to pathways, both because the data is very noisy (e.g., due to cross hybridization or low mRNA levels), and because similarity of expression profiles is only a weak indicator for the fact that two genes participate in the same pathway. Conversely, we can try to detect pathways by looking for groups of genes that contain many pairs of interacting genes. Once again, the reliability of these methods is low, both because the data is noisy, and because many gene products interact even when they are not part of the same pathway. For example, in the DIP binary interaction database (Xenarios *et al.*, 2000), 3527 genes form a single huge connected component.

We propose an approach that combines both types of data within a single probabilistic model, based on the framework of probabilistic graphical models (Pearl, 1988). Our approach aims to detect groups of genes that are co-expressed, and whose products interact in the protein interaction data. Specifically, we define a probabilistic model where genes are partitioned into "pathways". The likelihood of the data is higher when genes in the same pathway have the same expression profile; it is also higher when genes that interact are in the same pathway. The outline of our method is shown in Fig. 1. Starting from an input gene expression and protein interaction data, we first cluster the expression

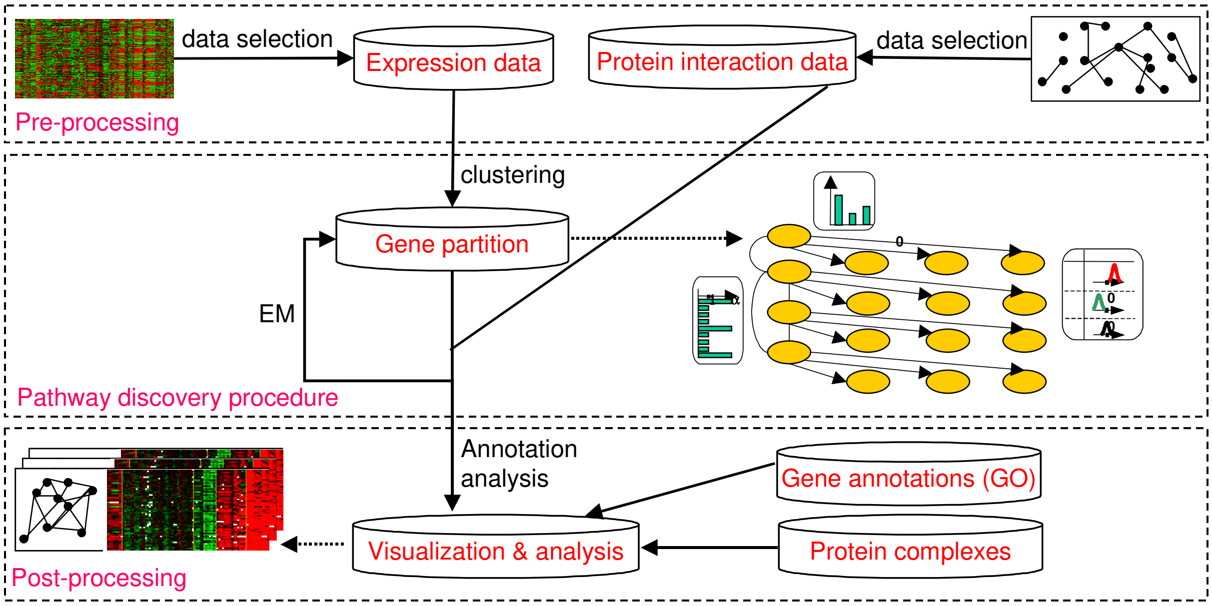


Fig. 1. Schematic flow diagram of our proposed method. The pre-processing step includes selecting the input gene expression and protein interaction data. The model is then trained using EM until convergence, and the resulting assignments of genes to pathways are then analyzed.

data, and create one cluster, or *pathway*, from each of the resulting clusters. These clusters serve to initialize the probabilistic model. The model is then trained to maximize the likelihood of the data, using the expectation maximization (EM) algorithm (Dempster *et al.*, 1977). Finally, we evaluate the biological performance of the model using external data sources that were not given as input to the model.

We evaluated the ability of our method to extract pathways from two different datasets of gene expression measurements combined with one binary protein interaction dataset. A comparison of our method to methods that use either the expression data or the protein interaction data alone shows that our inferred pathways correspond much better to known functional groups and protein complexes, both of which were not given as input to any of the methods.

PROBABILISTIC MODEL

In this section, we present our unified probabilistic model over gene expression and protein interaction data. Our model, which is based on the framework of *relational Markov networks* (Taskar *et al.*, 2002), defines a distribution over a set of genes $\mathbf{G} = \{g_1, \ldots, g_n\}$. We assume that each gene g belongs to precisely one of k pathways, denoted $g.C \in \{1, \ldots, k\}$. The variables $g_i.C$ are latent (or hidden) variables, and determining their values is one of the main goals of our algorithm. The model then has two components: one that models the expression data, and another that models the protein interaction data. The two easily combine to provide a single unified model.

Gene Expression Model. We use the simple *Naive Bayes* model for the gene expression component. In this approach, the instances are divided into disjoint classes, each of which is associated with a distribution over the attributes of the instances. The attributes are assumed to be conditionally independent given the class. Although this independence assumption is often unrealistic, this model has proven to be robust and effective for clustering in many applications (Cheeseman and Stutz, 1995; Duda *et al.*, 2000).

In our setting, the instances are the genes $g_1, \ldots, g_n$, and the class of g_i is simply the pathway to which it belongs. Each instance g has m continuous-valued attributes $g.\mathbf{E} = \{g.E_1, \ldots, g.E_m\}$, where $g.E_j$ represents the mRNA expression level measured for the gene in experiment j. The naive Bayes model defines a distribution:

$$P(g.C, g.E_1, \ldots, g.E_m) = P(g.C) \prod_{j=1}^{m} P(g.E_j \mid g.C).$$

The random variable $g.C$ is distributed as a multinomial distribution, parameterized by the vector $\theta_{\mathbf{C}} = \{\theta_{\mathbf{1}}, \ldots, \theta_{\mathbf{k}}\}$, where $P(g.C = p) = \theta_p$; thus, each $0 \leq \theta_p \leq 1$ and $\sum_{p=1}^{k} \theta_p = 1$. We model each *conditional probability distribution (CPD)*

$P(g.E_j \mid g.C = p)$ using the Gaussian distribution $\mathcal{N}(\mu_{pj}, \sigma^2_{pj})$. An illustration of the Naive Bayes model is given in Fig. 2(a). In a pure Naive Bayes model, the data consists of n instances (genes), each of which is sampled from this distribution.

Protein Interaction Model. Our probabilistic model for protein interaction data is based on our assumption that interacting proteins are more likely to be in the same pathway. Thus, if we observe a protein-protein interaction between the protein products of two genes, the genes are likely to belong to the same pathway. To model this assumption, we use the framework of *Markov networks* or *Markov random fields* (Kindeman and Snell, 1990), very common in statistical physics, e.g., to represent correlations between spins of neighboring electrons.

For our purposes, it suffices to define *binary Markov networks*. Let $\mathcal{V} = \{V_1, \ldots, V_n\}$ be a set of discrete random variables. A binary Markov network over $\mathcal{V}$ defines a joint distribution $P(\mathcal{V})$ as follows. The network is defined via an undirected graph whose nodes correspond to variables in $\mathcal{V}$ and whose edges $\mathcal{E}$ represent direct probabilistic interaction between those variables. Each variable V_i is associated with a *potential* $\phi_i(V_i)$. Each edge $[V_i—V_j]$ is associated with a non-negative *compatibility potential* $\phi_{i,j}(V_i, V_j)$. The joint distribution is then defined as

$$P(V_1, \ldots, V_n) = \frac{1}{Z} \prod_{i=1}^{n} \phi_i(V_i) \prod_{[V_i—V_j] \in \mathcal{E}} \phi_{i,j}(V_i, V_j),$$

where Z is a normalizing constant defined so as to make the distribution sum to 1. Intuitively, $\phi_i(V_i)$ encodes how likely the different values of V_i are, ignoring interactions between the variables. For an assignment v_i, v_j to V_i, V_j, the value $\phi_{i,j}(v_i, v_j)$ specifies how "compatible" the assignment v_i, v_j is: the higher the value, the more likely this pair of values is to appear together.

In the protein interaction setting, as in the work of Taskar *et al.* (2002), the variables are the classes of the instances in the data, and the edges are defined by relationships between them. Furthermore, parameters are shared across instances, so that we only have potentials $\phi_1(V_i)$ and $\phi_2(V_i, V_j)$. In our context, the variables $\mathcal{V}$ are the pathway assignments $g_1.C, \ldots, g_n.C$ of the genes in $\mathbf{G}$, and the edges correspond to protein-protein interactions observed in our data set. Intuitively, an edge between g_i and g_j captures our basic intuition that, if g_i and g_j interact, they are more likely to be in the same pathway. Thus, we define the compatibility potential $\phi_2(g_i.C = p, g_j.C = q)$ such that the compatibility value for $p = q$ is greater than the value for $p \neq q$. Since we do not assume any patterns over the distribution of interactions, we set all entries in which $p = q$ to the same value. Similarly, all entries in which $p \neq q$ are set to the same value. Due to the normalization of the distribution, what matters is only the relative magnitude of these two values. Thus, we can parameterize the interaction model using a single parameter, α, such that for all $[g_i—g_j] \in \mathcal{E}$:

$$\phi_2(g_i.C = p, g_j.C = q) = \begin{cases} \alpha & p = q \\ 1 & \text{otherwise} \end{cases} \quad (1)$$

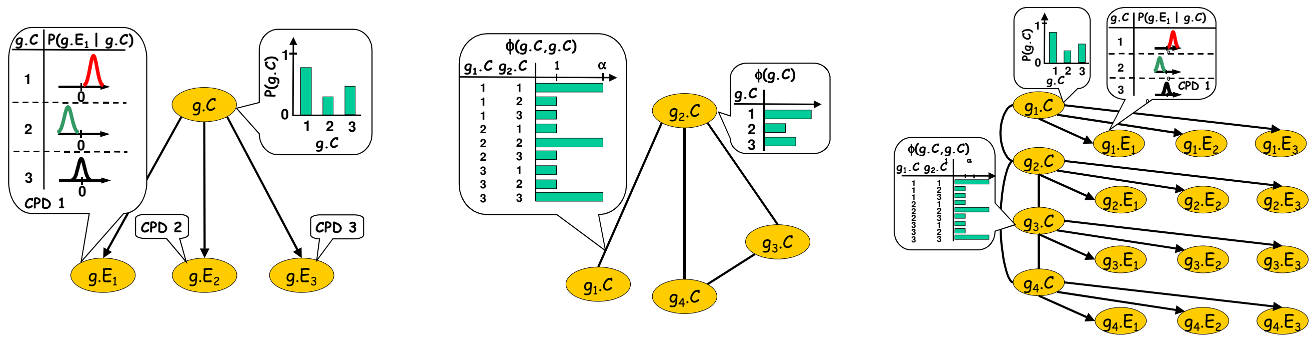
We require that $\alpha \geq 1$. Note that when $\alpha = 1$, these potentials have no effect on the distribution defined. The larger the value of α, the greater the induced distribution will be peaked around assignments in which genes that interact also belong to the same pathway. A simple example of the protein interaction model is given in Fig. 2(b).

Given a database of protein interactions which defines the set of edges $\mathcal{E}$, and the parameterizations for ϕ_1 and ϕ_2, the Markov network defines a joint distribution $P(g_1.C, \ldots, g_n.C)$ over assignments of genes to pathways. It is important to note that the assignments of different genes to pathways are *not* independent in this distribution: The model deliberately correlates the assignments of related genes. As a consequence, we cannot compute separately the pathway assignment of a single gene, and are forced to consider the distribution over the gene set as a whole. Unfortunately, this distribution is over an exponentially large space, so manipulating it exactly is intractable. We address this issue in the next section.

We note that the model we propose for protein interaction cannot stand by itself, as the assignment that maximizes the joint likelihood is degenerate: all genes are assigned to the same pathway. However, in our unified model that combines the protein interaction model with the expression model, this degenerate assignment is no longer the most likely.

Unified model. Our unified model integrates the models of the two subsections above. This combination can be performed very naturally, using the pathway variables $g_i.C$, that are common to both models. The distribution $P(g_i.C)$ used in the expression model can be used as the potential $\phi_1(g_i.C)$ in the interaction model. The remaining parameters — $P(g_i.E_j \mid g_i.C)$ in the expression model and $\phi_2(g_i.C, g_j.C)$ in the interaction model — do not conflict and can be placed in the same model.

The combined model is thus a partially-directed graphical model, with $m + 1$ random variables for each gene g_i: the pathway assignment $g_i.C$, and the expression values $g_i.E_1, \ldots, g_i.E_m$. The variable $g_i.C$ is associated with a multinomial distribution with parameters $\theta_{\mathbf{C}} = \{\theta_{\mathbf{C_1}}, \ldots, \theta_{\mathbf{C_k}}\}$. The CPD $P(g.E_j \mid g.C = p)$ is a Gaussian distribution $\mathcal{N}(\mu_{pj}, \sigma^2_{pj})$. Finally, each pair of genes g_i, g_j that interact are connected by an undirected edge, and associated with a compatibility



(a) (b) (c)

Fig. 2. (a) Naive Bayes model over 3 classes, for an expression data set with 3 expression measurements for each gene. A multinomial distribution is associated with $g.C$ (shown as a histogram). For each class $g.C$, each experiment is associated with a Gaussian CPD (shown in CPD 1). (b) Protein interaction model for a dataset with 4 genes in which the interactions are between: g_1 and g_2; g_2 and g_3; g_2 and g_4; and g_3 and g_4. Shown is the resulting Markov network, with its two types of potentials: $\phi_i(g_i.C)$ and $\phi_e(g_i.C, g_j.C)$. (c) Resulting unified partially-directed model.

potential $\phi_2(g_i.C, g_j.C)$, parameterized by a single α parameter as in Equation 1.

A simple example of this combined model is given in Fig. 2(c). The resulting combined model defines a joint distribution over the entire set of random variables, as follows:

$$P(\mathbf{G}.C, \mathbf{G}.\mathbf{E} \mid \mathcal{E}) = \frac{1}{Z}\left(\prod_{i=1}^{n} P(g_i.C) \prod_{j=1}^{m} P(g_i.E_j \mid g_i.C)\right) \cdot \left(\prod_{[g_i - g_j] \in \mathcal{E}} \phi_2(g_i.C, g_j.C)\right) \quad (2)$$

where Z is a normalizing constant that ensures that P sums to 1, and $\mathcal{E}$ represents all binary interactions that exist between genes in our data.

LEARNING THE MODEL

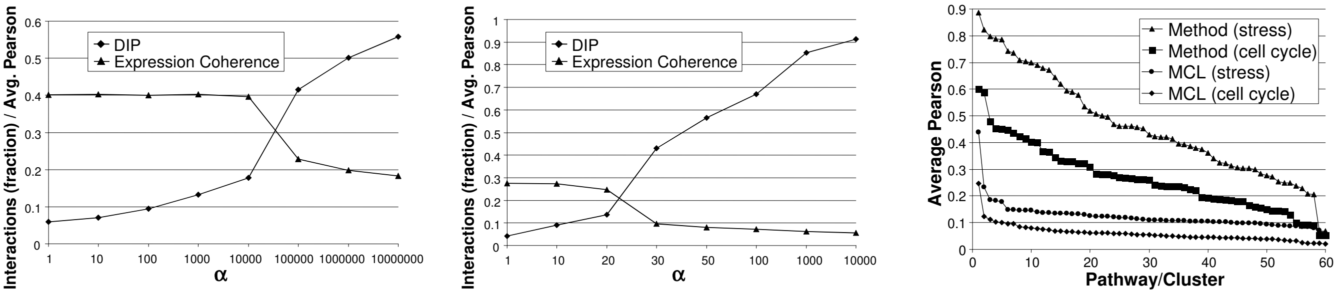
In the previous section we presented our unified model for expression and interaction data. In this section, we show how the parameters of this model are learned from data. Importantly, the pathway assignment variables $g_i.C$ are hidden, and are learned from data at the same time as the parameters.

Let $\mathbf{G}$ be a set of genes, and assume that we are given a dataset $\mathcal{D}$ that contains: for each gene g_i, an expression profile $g_i.\mathbf{E}$; and a set of binary interactions $\mathcal{E}$ between pairs of genes g_i, g_j. Our goal is to learn the model parameters Θ, which consist of: the multinomial θ over pathway assignments, and the means and standard deviations of each of the k Gaussian distributions associated with each of the $k \cdot m$ CPDs $P(g.E_i \mid g.C = p)$. Recall that the potentials over pairs of interacting genes are specified by a single α parameter. Here, we assume that α is given and do not consider estimating its value; we discuss the choice of α in the results section.

A standard approach is to find the *maximum likelihood* (ML) parameters: the parameters $\hat{\Theta}$ that maximize the likelihood $P(\mathcal{D} \mid \Theta)$. If we had a complete assignment to all the pathway variables, $\mathbf{G}.C$, then the likelihood function would have a unique global maximum, and the ML parameters could be found easily by computing the appropriate *sufficient statistics*. Specifically, for the pathway variables, the sufficient statistics are simply the frequencies of the different pathways: N_p = the number of genes assigned to pathway p. For the expression CPD $P(g.E_j \mid g.C = p)$, the sufficient statistics are the first and second moments of the empirical distribution: $\chi^1_{pj} = \sum_{\{i\,:\,g.C=p\}}^{n} g_i.e_j$ and $\chi^2_{pj} = \sum_{\{i\,:\,g.C=p\}}^{n} g_i.e_j^2$, where $g_i.e_j$ is the expression value of gene i in experiment j.

Unfortunately, our case, of incomplete data, is substantially more complex. In this case, the likelihood function has multiple local maxima, and no general method exists for finding the global maximum. The *Expectation Maximization (EM)* algorithm (Dempster *et al.*, 1977), provides an approach for finding a local maximum of the likelihood function. Starting from an initial guess $\Theta^{(0)}$ for the parameters, EM iterates two steps: an E-step and an M-step. The steps are iterated until convergence.

The **E-step** uses the current estimate of the parameters to compute the distribution over the hidden variables given the observed data. In our case, we compute $P(\mathbf{G}.C \mid \mathcal{D}, \Theta^{(t-1)})$. To compute this distribution, we must run inference over the entire partially-directed graphical model defined in Equation 2 and illustrated in Fig. 2(c). As we discussed, the interactions correlate the random variables associated with the different genes. When the interaction network is nontrivial, exact inference over this model is intractable. Instead, we use *belief*



(a) (b) (c)

Fig. 3. (a,b) Performance as a function of the potential parameter α. For protein interactions: fraction of all interactions in DIP that are between genes in the same pathway. For expression coherence: average Pearson correlation between pairs of genes in the same pathway. $\alpha = 1$ corresponds to a standard expression clustering model. (a) stress data; (b) cell cycle data. (c) Comparison of expression coherence of MCL clusters (measured as average Pearson correlation between pairs of genes in the same cluster) to our inferred pathways, where clusters were sorted by expression coherence.

propagation (Pearl, 1988), an approximate inference algorithm which passes messages between neighboring nodes in the graph. This algorithm has been shown to perform effectively for models of this type (Segal et al., 2001; Taskar *et al.*, 2002).

Using this posterior distribution, the **M-step** re-estimates the model parameters using the *expected* sufficient statistics, where the expectation is taken relative to this posterior. Letting $q(g,p) = P(g.C = p \mid D, \Theta^{(t-1)})$, then the expected sufficient statistics for the multinomial are simply $\overline{N}_p = \sum_{g \in \mathbf{G}} q(g,p)$. For the Gaussian CPD $P(g.E_j \mid g.C = p)$, the expected sufficient statistics are $\overline{\chi^1}_{pj} = \sum_{g \in \mathbf{G}} q(g,p) \cdot g.e_j$, and $\overline{\chi^2}_{pj} = \sum_{g \in \mathbf{G}} q(g,p) \cdot g.e_j^2$. We can then re-estimate the parameters by maximizing the likelihood with respect to the expected sufficient statistics:

$$\theta_p = \frac{\overline{N}_p}{\sum_{p'=1}^{k} \overline{N}_{p'}} \quad ; \quad \mu_{pj} = \frac{\overline{\chi^1}_{pj}}{\overline{N}_p} \quad ; \quad \sigma_{pj}^2 = \frac{\overline{\chi^2}_{pj}}{\overline{N}_p} - \mu_{pj}^2.$$

RESULTS

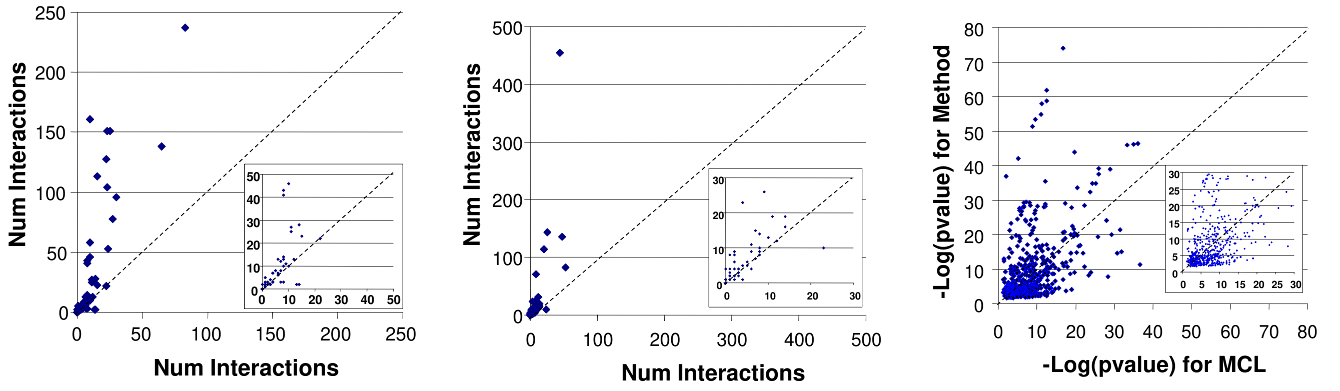
Model Learning. We evaluated our method on two *S. Cerevisiae* gene expression datasets, one consisting of 173 microarrays, measuring the responses to various stress conditions (Gasch *et al.*, 2000), and another consisting of 77 microarrays, measuring expression during cell cycle (Spellman *et al.*, 1998). For the protein interaction data, we used the DIP dataset (Xenarios *et al.*, 2000), consisting of 10705 *S. Cerevisiae* binary protein interactions. We selected only genes for which expression data was available from at least one of the datasets and that participated in binary interactions with at least one other gene in DIP. This resulted in a gene list consisting of 3589 genes, which we use in the experiments described below. We note that 3527 genes form a connected component in the interaction graph induced by DIP.

We applied our method to each of the expression datasets separately, combining each with the DIP interaction dataset. We trained each model using EM, as described in the previous section , fixing the number of pathways to be learned to be 60. A successful application of EM requires some reasonable initialization to the model parameters. To initialize the model, we applied the probabilistic hierarchical clustering algorithm of (Segal et al., 2001) to each expression dataset, resulting in a partition of genes into 60 clusters. We use this assignment to provide temporary values for the pathway variables $g.C$, and compute maximum likelihood estimates for the parameters relative to that assignment. These parameters form the starting point for EM, which was then run to convergence.

To complete the model parameterization, we need to specify α, the parameter used in Equation 1 to represent the strength of the preference towards assigning interacting genes to the same pathway. We experimented with a range of values for α for both data sets, measuring both the number of interactions in each pathway and the coherence of the pathways with respect to the expression profiles. We evaluated the expression coherence of a pathway as the average Pearson correlation between every pair of genes that were assigned to the pathway.[†]

As expected, increasing α results in a larger number of interactions among genes in the same pathway (see Fig. 3(a,b)). More surprisingly, for $1 \leq \alpha \leq 10000$ and $1 \leq \alpha \leq 20$ in the stress and cell cycle models, respectively, the quality of the gene expression patterns of each pathway were identical; this is surprising, since $\alpha = 1$ is equivalent to completely ignoring the interaction data and is thus the same as a standard clustering model which tries only to optimize the expression score.

[†]The Pearson correlation between two vectors $g_i.\mathbf{E}, g_j.\mathbf{E}$ is: *Pearson*$(g_i.\mathbf{E}, g_j.\mathbf{E}) = \frac{1}{m}\sum_{l=1}^{m} \frac{(g_i.E_l - \mu_i)}{\sigma_i} \frac{(g_j.E_l - \mu_j)}{\sigma_j}$, where μ_i, σ_i are the mean and standard deviation, respectively, of the entries in $g_i.\mathbf{E}$.



(a) (b) (c)

Fig. 4. (a),(b) Number of interactions between genes in pathways (y-axis) and in expression clustering (x-axis): (a) stress model; (b) cell cycle model. (c) Comparison of functional coherence using GO: Negative log *p*-value obtained for best pathway for each GO category (y-axis) versus negative log *p*-value for corresponding MCL cluster (x-axis) .

Thus, when using $\alpha = 10000$ and $\alpha = 20$ for the stress and cell cycle models, respectively, our method results in an organization into pathways that are much more consistent with the interaction data compared to an expression clustering model, while not sacrificing the gene expression quality. Consequently, we chose these settings for α. We note that the significant decrease in the expression score for higher values of α is due to the formation of a single large pathway, resulting from the domination of a high α value over the expression component of the model.

For our chosen values of α, we verified that the improved interaction consistency is not a result of a small number of pathways with dense interactions. We counted the number of interactions between genes in the same pathway separately for each pathway, and compared this to the corresponding cluster from which this pathway was initialized (as described above). Fig. 4(a,b) shows that the improvement is indeed distributed among many pathways.

We also compared our results to a method that uses only the interaction data. We used the graph clustering *Markov Cluster Algorithm* (MCL) of Enright *et al.* (2002). We applied MCL to the DIP data, resulting in 905 clusters. To allow for a comparison with our models, we reduced the number of clusters to 60, by iteratively merging the two clusters whose resulting merged cluster had the lowest probability of observing its number of interactions by chance (computed using a binomial distribution as the null model), until we were left with 60 clusters. As expected, since MCL only tries to optimize the interaction score, the total number of interactions between genes assigned to the same pathway was greater for MCL (5261 such interactions) compared to our method (1913 interactions). However, the expression data does not support the organization of MCL, as can be seen in Fig. 3(c) which compares the expression score of the MCL clusters to our pathways.

Evaluation. We evaluated our learned models relative to a variety of external data sources, that were not used in learning the models. The visualization and statistical analysis of the results were performed using GeneXPress (available from `http://GeneXPress.stanford.edu`), a generic cluster analysis and visualization software that we developed. We evaluated the models along several criteria: prediction of held-out interactions, coherence of pathways according to functional annotations, and coverage of protein complexes.

Recall that our definition of a pathway is a set of genes whose protein products are more likely to interact. We would like to test the quality of our pathways relative to this criterion. However, it would not be surprising if our pathways contained a large number of interactions that are part of the input to our algorithm: the algorithm explicitly tries to put interacting genes in the same pathway. We therefore evaluated predictiveness of interactions by hiding some interactions from the learning algorithm, and testing how many of these were predicted based on the remaining data. Thus, we used 5-fold cross-validation, randomly partitioning the DIP data into 5 equally sized partitions, and trained 5 different models, each on a different subset of 4 of the interaction partitions. We then tested the predictiveness of each model on the held out data by counting the total number of held out interactions between genes assigned to the same pathway. This number was averaged over pathways, and over the 5 models learned. For the stress data set, there were 222.4 ± 13.2 such interactions for our method, compared to 126 ± 4.1 for expression clustering, and 383.2 ± 29.1 for MCL. It is not surprising that MCL performs better along this metric, as it optimizes only for interaction density, whereas our approach tries to capture both interactions and expression

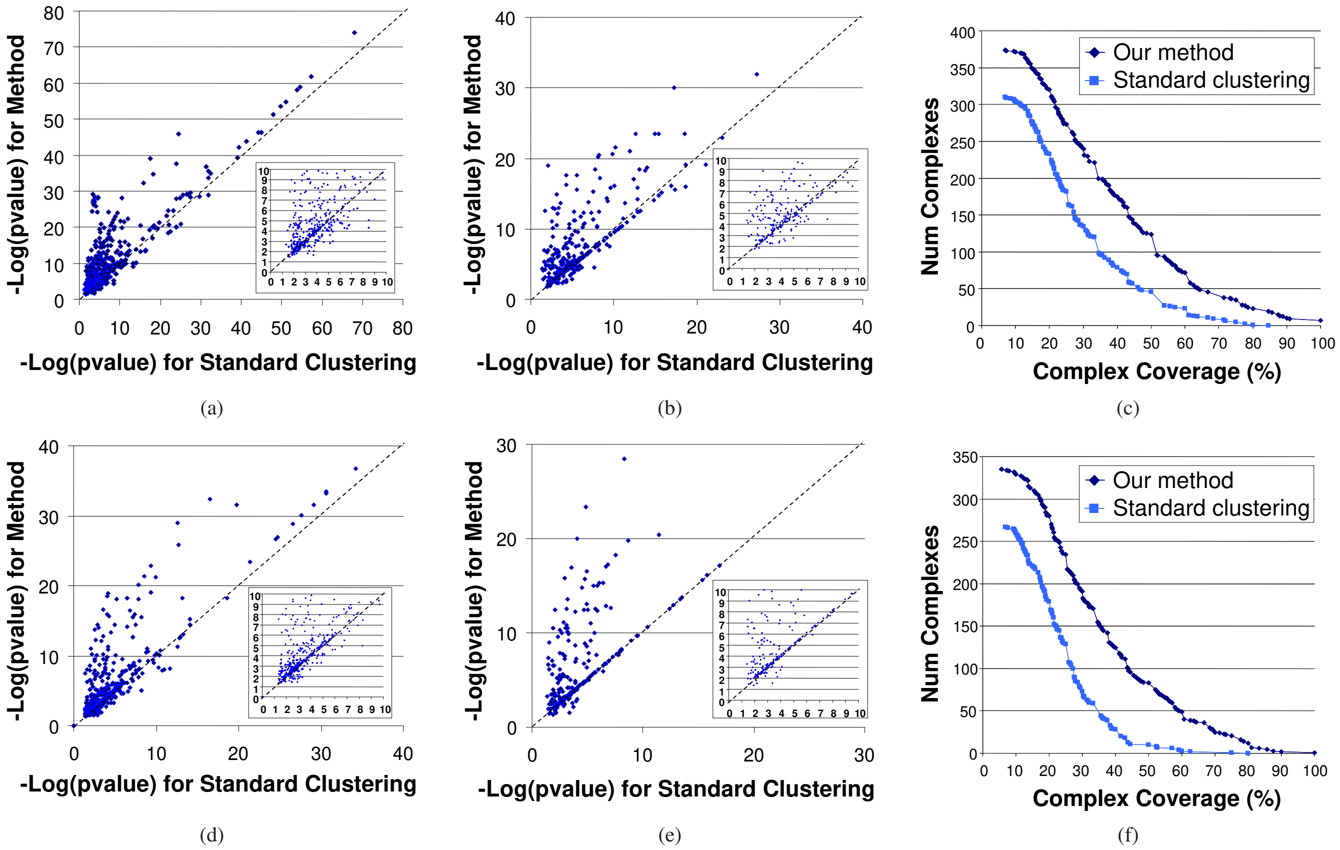


Fig. 5. Comparison of standard clustering and the proposed method: (a)–(c) stress model; (d)–(f) cell cycle model. (a,d) Comparison of functional coherence using GO: Negative log p-value obtained for best pathway for each GO category (y-axis) versus negative log p-value for corresponding cluster (x-axis). (b,e) Comparison of correspondence to protein complexes: Negative log p-value obtained for best pathway for each complex (y-axis) versus negative log p-value for corresponding cluster (x-axis). (c,f) The degree to which entire complexes appear together in the inferred pathways: The number of complexes for which a fraction q or more of their member genes appear in the same pathway or cluster, for all possible values of q.

coherence.

Both expression coherence and interaction density are only weak indicators for a pathway. To analyze the biological coherence of the inferred pathways, we computed their enrichment for annotations from the GO hierarchy (Ashburner *et al.*, 2000). We used the *S. cerevisiae* GO associations from SGD (Cherry *et al.*, 1998) to associate each gene with the processes it participates in, and removed all annotations associated with less than 5 genes. This resulted in 537 categories. For each pathway and each annotation, we calculated the fraction of genes in the pathway associated with that annotation and used the hypergeometric distribution to calculate a p-value for this fraction. We performed a Bonferroni correction for multiple independent hypotheses and took p-value $< 0.05/N$ ($N = 537$) to be significant.

The models learned spanned a wide variety of functional categories and pathways, including energy, respiratory, translational, transport, transcriptional regulation, cell organization, DNA replication, and protein degradation pathways. Furthermore, a comparison of the best p-values learned for each category (Fig. 5(a,d)) shows that the pathways learned by our model are more coherent functionally than the clusters learned by standard clustering. Not surprisingly, the highly coherent categories were detected by both methods, although the coherence was still marginally better using our method. More interesting, for the less clear cases (negative log p-value $<$ 10), a large number of categories were much more significantly enriched using our method (see zoomed plot in figures).

We did a similar comparison with the clusters resulting from the graph clustering method, MCL, which only optimizes the protein interaction data (see Fig. 4(c)). In the less clear range (negative log p-value $<$ 10) the results were comparable, but there were many more highly coherent categories (negative log p-value $>$ 20) enriched in our pathways (73 such categories) compared to MCL (32 such categories). Among the functional groups much more enriched in our pathways were categories related to translation (e.g., ribosome), protein degradation (e.g., proteasome), transcription (e.g., transcription from Pol I) and DNA replication. Genes in these categories interact with many genes from other categories in DIP and are thus hard to isolate using MCL. However, these categories are often co-expressed, which explains the success of our method in isolating them.

The components of many pathways are protein complexes. Thus, a good pathway model should assign the member genes of many of these complexes to the same pathway. We tested whether our models exhibited this property using the recent experimental assays of Gavin *et al.* (2002) and Ho *et al.* (2002), which assayed the members of 590 and 493 protein complexes, respectively.

We first measured the overlap between the protein complex data and DIP to verify that we can indeed treat the complex data as an independent data source. To do so, we converted the complex data into binary interactions, by creating a binary interaction between every pair of genes that are in a complex together, and measured the overlap with the DIP interactions. Only 2633 are shared, out of the 48751 and 10705 binary interactions in the complex and DIP data, respectively. Given that the complex assays are different in nature from the DIP binary interactions, and given this small overlap, we concluded that we could use the complex data as an independent measure of performance.

To analyze whether a pathway is significantly enriched for protein complexes, we associated each gene with the complexes to which it belongs in the complex experimental datasets. For each pathway and each complex, we computed the enrichment p-value , similar to the computation done for the GO annotation enrichment (Bonferroni corrected). From a total of 640 complexes, 374 were significantly enriched in at least one of the inferred pathways. Fig. 5(b,e) compares the complex enrichment between our model and standard clustering, indicating a much higher enrichment in our models. We also tested the degree to which entire complexes appear in the same pathway, by counting the number of complexes for which a fraction q or higher of their member genes appear in the same pathway. Fig. 5(c,f) shows the results for varying values of q. For example, for the stress model, there are 124 complexes for which 50% or more of their members appeared in the same pathway, compared to only 46 such complexes in the standard clustering model. For the cell cycle model, there are 83 complexes at 50% compared to only 10 in the standard clustering model.

Detailed inspection of the inferred pathways revealed many cases in which our method isolated known pathways from the dense web of DIP interactions, and potentially also identified novel members of the pathway. Such was the case for pathway 1, whose largest connected component of DIP interactions had 12 genes, 10 of which are members of the cytoplasmic exosome (RNase complex), required for the 3' processing of pre-rRNAs to mature rRNAs. These 10 genes are all the known members of this complex, so that our approach captured the complex (as it is currently known) in its entirety. One of the two remaining genes was YHR081W, an uncharacterized protein. As YHR081W interacts with 2 proteins in the pathway, and as its expression profile is highly similar to that of the other 11 genes in the component, this may be a potential novel discovery of a new member or related member of the RNase. We note that there were 38 additional immediate neighbors of these 12 genes in the DIP interaction graph that our method did not assign to pathway 1, since their expression profiles were different than those of these 12 genes.

We checked whether methods that analyze only the expression data or only the interaction data were also successful in isolating the RNase. The expression clustering method assigned only 4 of the 10 cytoplasmic exosome to the same cluster (see Fig. 6). As there are many interactions between the 10 RNase genes, MCL also assigned all 10 to the same cluster. However, the connected component included 114 additional genes that are not known to be related to the cytoplasmic exosome. Interestingly, MCL also included the same uncharacterized ORF YHR081W as our method.

As another example, our method identified 19 of the 25 genes that are annotated in GO as nucleus import genes, as part of a connected component of 55 genes in pathway 9. In contrast, the expression-based clustering method only had 4 of these genes in a cluster, and MCL grouped only 7 of these genes as part of a connected component of 97 genes.

DISCUSSION AND CONCLUSIONS

We presented a unified probabilistic model over both gene expression and protein interaction data, that searches for *pathways* — sets of genes that are both co-expressed and whose protein products interact. We showed that our method discovers groups of genes that correspond better to functional groups, and contain entire protein complexes, properties that one would expect of a pathway.

Our models currently constrain each gene to be in exactly one pathway. This is clearly a limitation, since in different conditions genes participate in different cellular processes. Recently, several approaches have been proposed that discover condition-specific groups from gene expression (Ihmels *et al.*, 2002; Segal et al., 2001). The discovery of condition-specific groupings would have great potential in the context of protein interactions, as it may allow us to identify which interactions are active under which conditions.

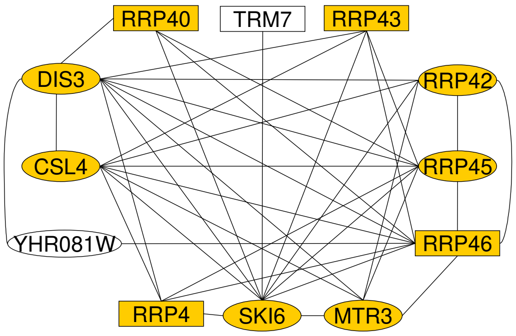


Fig. 6. Largest connected component, consisting of 12 genes, for one of the pathways identified. Nodes represent genes. Links between nodes indicate interactions in DIP. Square nodes (5) correspond to genes that were also in the pathway using standard clustering. Oval nodes (7) are genes that were only in the pathway using our method. Filled nodes (10) are all the 10 cytoplasmic exosome (RNase complex) annotated genes in GO. Unfilled nodes are either annotated differently (TRM7) or unknown (YHR081W).

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation, grant ACI-0082554. Eran Segal was also supported by a Stanford Graduate Fellowship (SGF).

REFERENCES

Alon,U., N.Barkai, D.A.Notterman, K.Gish, S.Ybarra, D.Mack, and A.J.Levine (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS 96*(12), 6745–50.

Ashburner,M., C.A.Ball, J.A.Blake, D.Botstein, H.Butler, J.M.Cherry, A.P.Davis, K.Dolinski, S.S.Dwight, J.T.Eppig, M.A.Harris, D.P.Hill, L.Issel-Tarver, A.Kasarskis, S.Lewis, J.C.Matese, J.E.Richardson, M.Ringwald, G.M.Rubin, and G.Sherlock (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics 25*, 25–9.

Cheeseman,P. and J. Stutz (1995) Bayesian classification (AutoClass): Theory and results. In *Advances in Knowledge Discovery and Data Mining*, pp. 153–180. Menlo Park, CA: AAAI Press.

Cherry,J. M. *et al.* (1998) Sgd: Saccharomyces genome database. *Nucleic Acid Research 26*, 73–9.

Dempster,A. P., N. M. Laird, and D. B. Rubin (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B 39*, 1–39.

Duda,R. O., P. E. Hart, and D. G. Stork (2000) *Pattern Classification*. New York: John Wiley & Sons.

Eisen,M.B., P.T.Spellman, P.O.Brown, and D.Botstein (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS 95*, 14863–68.

Enright,A.J., S.Van Dongen, and C.A.Ouzounis (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acid Research 30*, 1575–84.

Gasch,A. P., P.T.Spellman, C.M.Kao, O.Carmel-Harel, M.B.Eisen, G.Storz, D.Botstein, and P.O.Brown (2000) Genomic expression program in the response of yeast cells to environmental changes. *Mol. Bio. Cell 11*, 4241–4257.

Gavin A.C., M.Bosche, R.Krause, P.Grandi, M.Marzioch, A.Bauer, J.Schultz, J.M.Rick, A.M.Michon, C.M.Cruciat, M.Remor, C.Hofert, M.Schelder, M.Brajenovic, H.Ruffner, A.Merino, K.Klein, M.Hudak, D.Dickson, T.Rudi, V.Gnau, A.Bauch, S.Bastuck, B.Huhse, C.Leutwein, M.A.Heurtier, R.R.Copley, A.Edelmann, E.Querfurth, V.Rybin, G.Drewes, M.Raida, T.Bouwmeester, P.Bork, B.Seraphin, B.Kuster, G.Neubauer, G.Superti-Furga (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature 415*, 141–7.

Ge,H., Z. Liu, G. Church, and M. Vidal (2001) Correlation between transcriptome and interactome mapping data from s. cerivisia. *Nat. Gen. 29*, 482–6.

Ho Y., A.Gruhler, A.Heilbut, G.D.Bader, L.Moore, S.L.Adams, A.Millar, P.Taylor, K.Bennett, K.Boutilier, L.Yang, C.Wolting, I.Donaldson, S.Schandorff, J.Shewnarane, M.Vo, J.Taggart, M.Goudreault, B.Muskat, C.Alfarano, D.Dewar, Z.Lin, K.Michalickova, A.R.Willems, H.Sassi, P.A.Nielsen, K.J.Rasmussen, J.R.Andersen, L.E.Johansen, L.H.Hansen, H.Jespersen, A.Podtelejnikov, E.Nielsen, J.Crawford, V.Poulsen, B.D.Sorensen, J.Matthiesen, R.C.Hendrickson, F.Gleeson, T.Pawson, M.F.Moran, D.Durocher, M.Mann, C.W.Hogue, D.Figeys, M.Tyers (2002) Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectometry. *Nature 415*, 180–3.

Ihmels,J., G.Friedlander, S.Bergmann, O.Sarig, Y.Ziv, and N.Barkai (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Gen. 4*, 370–7.

Jansen,R., D. Greenbaum, and M. Gerstein (2002) Relating whole genome expression data with protein protein interactions. *Genome Research 12*, 37–46.

Kindeman,R. and J. Snell (1990) *Markov random fields and their applications*. Amer. Math Soc.

Pearl,J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.

Segal,E., B. Taskar, A. Gasch, N. Friedman, and D. Koller (2001). Rich probabilistic models for gene expression. *Bioinformatics 17*(Suppl 1), S243–52.

Spellman,P.T., G.Sherlock, M.O.Zhang, V.R.Iyer, K.Anders, M.B.Eisen, P.O.Brown, D.Botstein, and B.Futcher (1998) Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell 9*(12), 3273–97.

Taskar,B., P. Abbeel, and D. Koller (2002) Discriminative probabilistic models for relational data. In *Proc. UAI*.

Xenarios,I., D.W.Rice, L.Salwinski, M.K.Baron, E.M.Marcotte, and D.Eisenberg (2000) Dip: The database of interacting proteins. *Nucleic Acid Research 28*, 289–91.