Research

Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering

Audrey P Gasch* and Michael B Eisen*†

Addresses: *Department of Genome Science, Life Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. †Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA.

Correspondence: Michael B Eisen. E-mail: mbeisen@lbl.gov

Published: 10 October 2002

Genome ***Biology*** 2002, **3(11)**:research0059.1–0059.22

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/11/research/0059

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 1 August 2002
Revised: 6 September 2002
Accepted: 11 September 2002

Abstract

Background: Organisms simplify the orchestration of gene expression by coregulating genes whose products function together in the cell. Many proteins serve different roles depending on the demands of the organism, and therefore the corresponding genes are often coexpressed with different groups of genes under different situations. This poses a challenge in analyzing whole-genome expression data, because many genes will be similarly expressed to multiple, distinct groups of genes. Because most commonly used analytical methods cannot appropriately represent these relationships, the connections between conditionally coregulated genes are often missed.

Results: We used a heuristically modified version of fuzzy k-means clustering to identify overlapping clusters of yeast genes based on published gene-expression data following the response of yeast cells to environmental changes. We have validated the method by identifying groups of functionally related and coregulated genes, and in the process we have uncovered new correlations between yeast genes and between the experimental conditions based on similarities in gene-expression patterns. To investigate the regulation of gene expression, we correlated the clusters with known transcription factor binding sites present in the genes' promoters. These results give insights into the mechanism of the regulation of gene expression in yeast cells responding to environmental changes.

Conclusions: Fuzzy k-means clustering is a useful analytical tool for extracting biological insights from gene-expression data. Our analysis presented here suggests that a prevalent theme in the regulation of yeast gene expression is the condition-specific coregulation of overlapping sets of genes.

Background

All organisms possess an essentially fixed repertoire of proteins determined by their genome sequence. They have evolved to survive varying internal and external environments by carefully controlling the abundance and activity of these proteins to suit their conditions. To simplify this task, genes whose products function together are often under common regulatory control such that they are coordinately expressed under the appropriate conditions. This property has been frequently exploited in the analysis of genome-wide expression data, as the experimental observation that a set of genes is coexpressed frequently implies that the genes share a biological function and are under common regulatory control [1]. Many proteins have multiple roles in the

cell, however, and act with distinct sets of cooperating proteins to fulfill each role. Their genes are therefore coexpressed with different groups of genes, each governed by a distinct regulatory mechanism, in response to the varying demands of the cell (Figure 1a). This complicates the analysis of expression data and calls for a more nuanced approach to data analysis.

The yeast *Saccharomyces cerevisiae* evolved in a niche in which the availability of nutrients and the conditions of growth vary constantly, and it possesses sophisticated mechanisms to choreograph the expression of its approximately 6,000 genes in order to thrive - or at least survive - in a wide range of environmental conditions. These responses are governed by a complex, condition-specific regulatory system that transduces information through the cell to the nucleus, where gene expression is adjusted accordingly. Many of the individual components of this regulatory system function under particular conditions and govern the expression of overlapping sets of gene targets, allowing a given gene to be coexpressed with different gene groups in response to different conditions (Figure 1a). As a consequence, the targets of each regulatory system often display similar expression patterns in response to one set of conditions but divergent patterns under other situations (Figure 1b). For example, the known targets of the oxidative stress-responsive transcription factor Yap1p are coordinately induced in response to conditions that inflict oxidative damage, but these genes are divergently expressed in response to other environmental changes (Figure 1c) [2]. Similarly, the known targets of other transcription factors in yeast (including Aft1p, Zap1p, Pho4p, Hac1p, Hsf1p, and others) are similarly expressed only in response to certain environments [2-6].

The complexity of the regulatory network that governs yeast gene expression complicates the analysis of whole-genome expression data. Because of the connection between gene-expression regulation and gene product function, computational analysis of expression data is used extensively to identify groups of similarly expressed genes. However, the central limitation of most of the commonly used algorithms is that they are unable to identify genes whose expression is similar to multiple, distinct gene groups, thereby masking the relationships between genes that are coregulated with different groups of genes in response to different conditions. Consider, for example, k-means clustering [7,8]. The k-means algorithm partitions genes into a defined set of discrete clusters, attempting to maximize the expression similarity of the genes in each cluster (Figure 2a). The algorithm is initiated by randomly partitioning the genes into *k* groups. Each group is then represented by a 'centroid' (the mean expression pattern of genes in the group), and the genes are repartitioned to the cluster whose centroid is most similar to their expression pattern. The partitioning process is iterated until the gene partitions are stable (or some other stopping criterion is met). The end result of the algorithm is a set of *k* clusters of similarly expressed genes. However, a key property of this algorithm (and many others like it) is that each gene is assigned to one and only one cluster, obscuring the relationships between conditionally coregulated genes such as those shown in Figure 1. This limitation is especially problematic when analyzing large gene-expression datasets that are collected over many experimental conditions, when many of the genes are likely to be similarly expressed with different groups in response to different subsets of the experiments. A number of methods have been developed to deal with complex relationships between objects [9-11]. Here, we explore the utility of one such method - fuzzy k-means clustering.

Fuzzy k-means clustering [12] facilitates the identification of overlapping groups of objects by allowing the objects to belong to more than one group. The essential difference between fuzzy k-means clustering and standard k-means clustering is the partitioning of genes into each group (Figure 2b). Rather than the hard partitioning of standard k-means clustering, where genes belong to only a single cluster, fuzzy k-means clustering considers each gene to be a member of every cluster, with a variable degree of 'membership'. Each gene has a total membership of 1.0 that is apportioned to clusters on the basis of the similarity between the gene's expression pattern and that of each cluster centroid. Genes whose expression patterns are very similar to a given centroid will be assigned a high membership in that cluster, whereas genes that bear little similarity to the centroid will have a low membership. Importantly, genes can be assigned significant memberships to more than one cluster, thus revealing genes whose expression is similar to multiple, distinct groups of genes.

We implemented a heuristic variant of fuzzy k-means clustering that incorporated principal component analysis (PCA) and hierarchical clustering to analyze published yeast genomic expression data that followed the response of cells to different environments. The method successfully identified clusters of functionally related genes and more comprehensive groups of known transcription factor targets in yeast. In the process of this analysis, we identified previously unrecognized similarities in the expression of yeast genes and uncovered correlations between the environmental conditions. We explored the regulation of gene expression by correlating the identified clusters with known regulatory elements present in the genes' promoters. These details implicate mechanisms that yeast cells use to orchestrate genomic expression programs in response to variable conditions.

Results

Fuzzy k-means clustering overview

We implemented a version of the fuzzy k-means algorithm, based on a description by Gath and Geva [13], in a C++ program called FuzzyK (available at [14]; see Materials and Methods for complete details). We altered the algorithm in

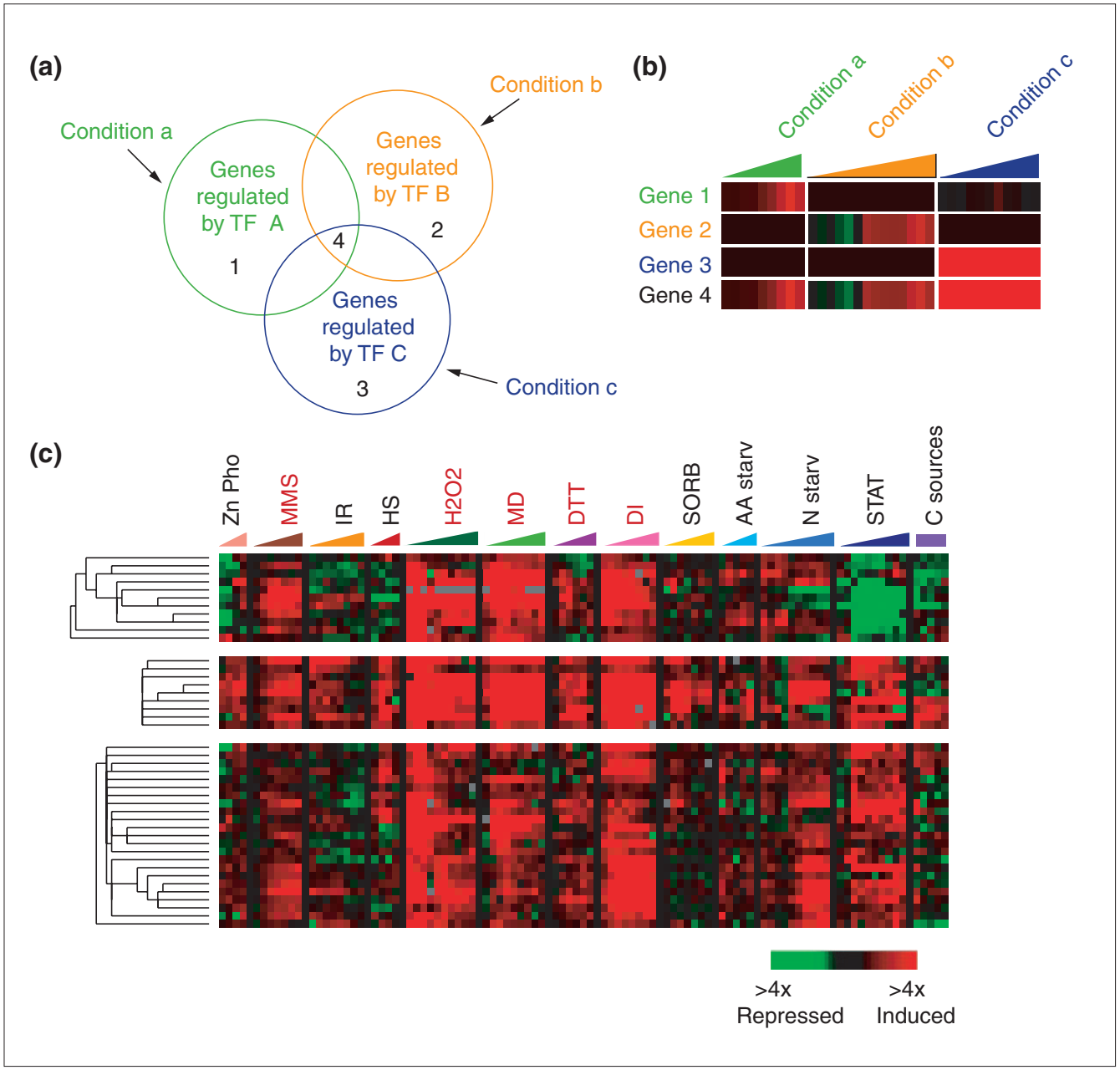
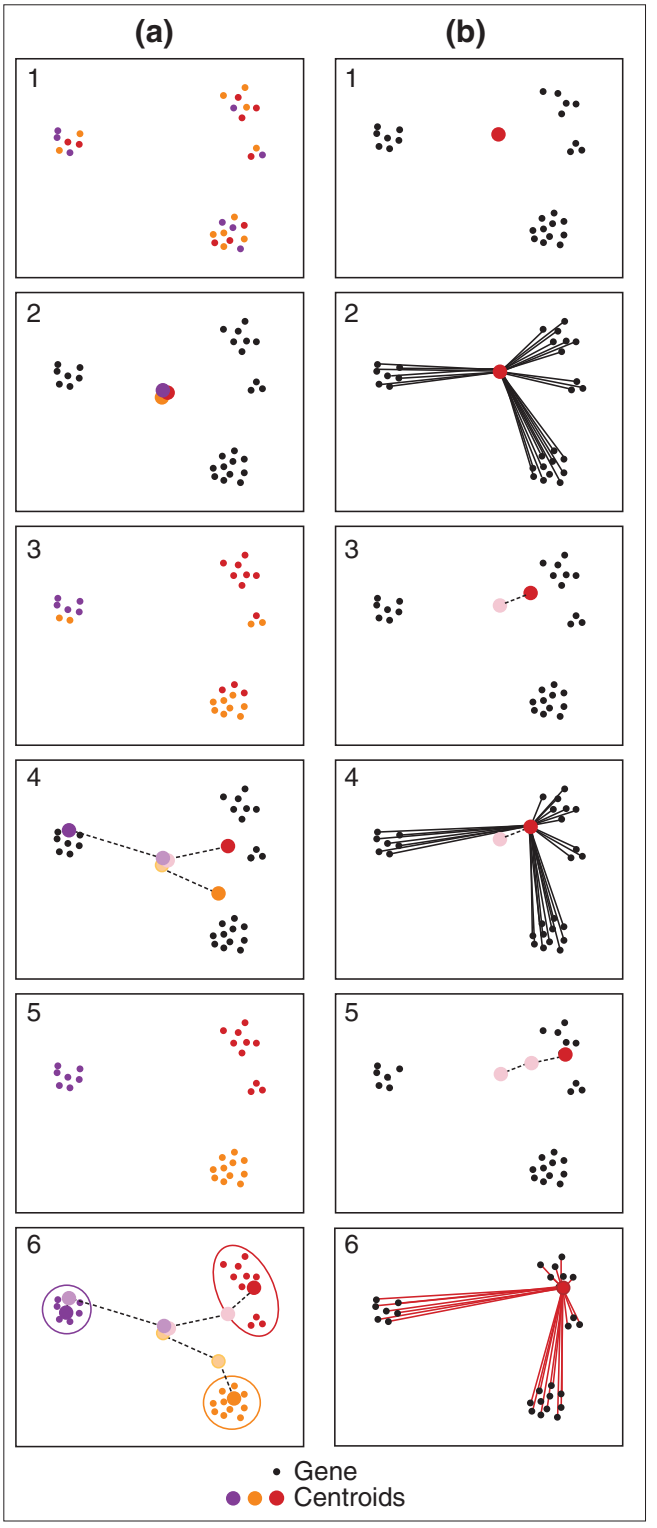


Figure 1
Many yeast genes are conditionally coregulated. **(a)** A Venn diagram representing hypothetical genes that are coregulated by transcription factor A (TF A) in response to condition a, transcription factor B (TF B) in response to condition b, or transcription factor C (TF C) in response to condition c. The regions of overlap in the diagram represent genes that are conditionally coregulated with each respective group of genes (for example, gene 4). **(b)** Hypothetical gene-expression patterns for four representative genes in groups from (a) show that the expression pattern for gene 4 has similarities to the expression patterns of each of the other genes. For this and other diagrams, gene-expression data are represented in a colorized, tabular format in which each row indicates the relative transcript abundance for a given gene, and each column represents the relative transcript abundance for many genes as measured in one experiment. A red square indicates that a gene was induced in response to the condition listed, a green square indicates that a gene was repressed under those conditions, a black square indicates that there was no detectable change in expression, and a gray square represents missing data. **(c)** The gene-expression patterns of around 40 of the 70 known Yap1p targets are shown, as the genes appear in the complete, hierarchically clustered dataset. Because these genes were coordinately induced in response to only subsets of the conditions shown here (labeled in red), the entire set of Yap1p targets was assigned to multiple hierarchical clusters, the largest of which are shown here. The remaining Yap1p targets were assigned to other hierarchical clusters and are not shown in this display. The colored triangles above the figure represent the microarray time courses that measured the changes in transcript abundance in response to zinc or phosphate limitation (Zn Pho), treatment with methylmethane sulfonate (MMS), ionizing radiation (IR), heat shock (HS), hydrogen peroxide (H_2O_2), menadione (MD), dithiothreitol (DTT), diamide (DI), sorbitol (SORB), amino-acid starvation (AA starv), nitrogen starvation (N starv), and progression into stationary phase (STAT). Steady-state gene expression was also measured for cells growing on alternative carbon sources (C sources), indicated by the purple rectangle. See text for references.

two fundamental ways: first, we performed three successive cycles of fuzzy k-means clustering, with the second and third rounds of clustering performed on subsets of the data. Second, because the random initialization commonly used in k-means clustering can have a profound impact on the results [15], we instead chose to initialize each clustering cycle by seeding prototype centroids with the eigen vectors identified by PCA of the respective dataset (see below). Here we present an overview of the algorithm, followed by a discussion of the parameter optimization.



The input of the program is a table of expression values, where each row represents a given gene's relative transcript abundance under the condition indicated in each column (Figure 3). The first round of clustering is initialized by defining $k/3$ prototype centroids (where k is the total number of clusters and 3 is the number of clustering cycles) as the most informative $k/3$ eigen vectors identified by PCA of the input dataset (see Materials and methods). The prototype centroids are refined in the subsequent steps: each gene is assigned a membership score to each of the prototype centroids, based on the Pearson correlation between the gene's expression pattern and the centroid in question. Each of the centroid patterns is then recalculated as the weighted mean of all of the gene-expression patterns in the dataset, where each gene's weight is proportionate to its membership in the corresponding cluster (see Materials and methods for details regarding the calculations). Genes that have a large membership to a given centroid will contribute more to the mean, and the new centroid will migrate in the direction of those genes. The process of calculating gene-centroid memberships and updating the centroids is iterated until the centroid patterns become fixed (or until the termination criterion is met, as described in Materials and methods).

After this initial round of fuzzy clustering, duplicate centroids (pairs whose Pearson correlation is greater than 0.9) are averaged, and genes with a greater than 0.7 correlation to any of the identified centroids are removed from the dataset (see Materials and methods). The fuzzy k-means clustering steps described above are repeated on this smaller

Figure 2
Comparison of k-means and fuzzy k-means clustering. Genes are represented as points in space, where genes that are similarly expressed are close together. **(a)** An overview of standard k-means clustering. (1) The process is initiated by randomly partitioning the genes (small circles) into three groups, indicated by the three colors. (2) The average expression profile of each group of genes is calculated as the centroid (large circles), and the genes are reassigned to the centroid to which they are closest. (4-6) Steps 2 and 3 are iterated until the centroids are stable, at which point the genes are assigned to the cluster to which they are most similar. **(b)** An overview of fuzzy k-means clustering. (1) The process is initiated by seeding each centroid with an eigen vector identified by PCA, as shown here for one centroid (large circle). (2) For a given centroid, the membership of each gene is calculated from the distance (or similarity) between each gene-expression pattern and the centroid. (3) A new centroid is calculated as the weighted average of all of the gene-expression patterns in the dataset, where each gene's weight is proportionate to its membership in the cluster. Genes that are closer to the centroid will contribute more to the cluster mean; therefore the centroid position migrates toward those genes. (4-6) The process is iterated until convergence, and the membership of each gene in each cluster is calculated (as shown here for one cluster).

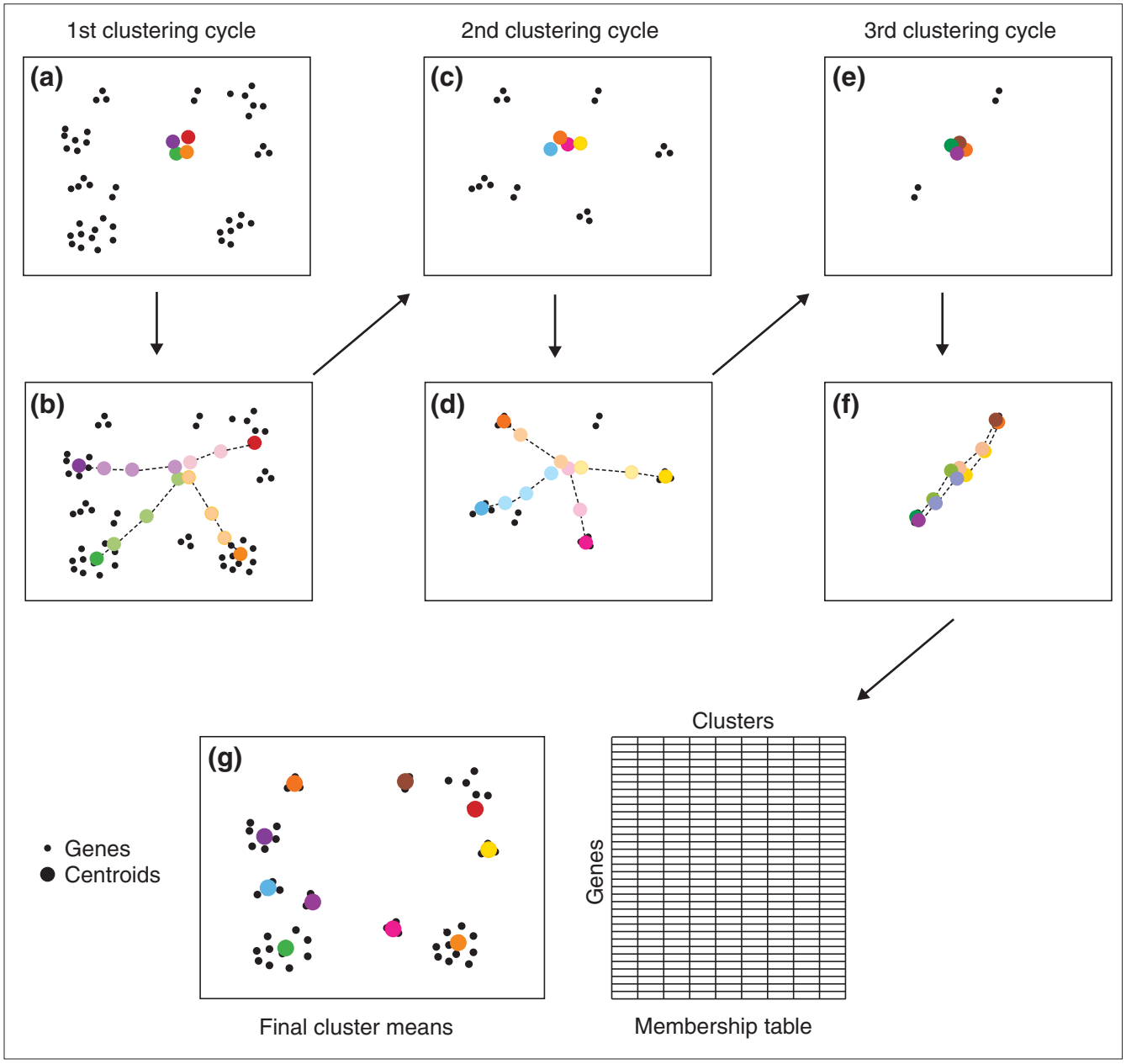


Figure 3
Overview of the FuzzyK method. Genes are represented as points in space, where genes that are similarly expressed are close together. **(a)** In the first fuzzy-clustering cycle, *k*/3 centroids are defined as the most informative *k*/3 eigen vectors identified by PCA of the input dataset (large colored circles). **(b)** The centroids are refined by iteratively calculating the gene-cluster memberships and updating the centroid positions until convergence (see Figure 2b). **(c,d)** Genes that are correlated >0.7 to the identified centroids are removed from the dataset, gene and array weights are recalculated, and the entire fuzzy k-means clustering process is repeated on the data subset for an additional *k*/3 clusters (see Materials and methods for details). **(e,f)** Steps c and d are repeated for a third round of fuzzy clustering. **(g)** The output of the algorithm is a list of unique centroids and a table of gene-cluster memberships.

dataset to identify patterns missed in the first clustering cycle, and the new centroids are added to the set identified in the first round. The process of averaging replicated centroids and selecting a data subset is repeated, and a third cycle of clustering is performed on the subset of genes with a correlation of less than 0.7 to any of the existing centroids. The newly identified centroids are combined with the previous sets, and replicate centroids are averaged.

In the final step of the program, the membership of each gene to each centroid is calculated. Thus, the output of the algorithm is twofold: the method presents a list of the

unique centroids identified in the fuzzy clustering cycles along with a matrix representing the final membership scores for each gene to each centroid. In this representation, each gene can be related to all the identified clusters through its membership value, allowing genes that belong significantly to multiple clusters to be realized. As a consequence, each cluster consists of a continuous list of all of the genes in the dataset, ranked according to decreasing membership.

The continuous clusters identified by fuzzy k-means clustering present a challenge in visualizing the clustering results. To this end, we have developed the program FuzzyExplorer, a PERL viewer based on the program GeneExplorer (developed by Christian Rees; C. Rees, P.O. Brown and D. Botstein, unpublished results). Using this software, the genes that belong significantly to each cluster can be identified and visualized by applying a membership cutoff: all genes whose membership is greater than the cutoff will be selected as part of the cluster and their gene-expression patterns will be displayed. Rather than define a single cutoff for each cluster, the visualization software applies a sliding membership cutoff to select the genes, allowing each cluster to be expanded or collapsed in terms of the number of genes selected. This flexibility allows the user to define the appropriate membership cutoff for each cluster. For example, at a very high membership cutoff, most of the genes in each cluster will have highly correlated expression patterns in all of the experiments and will be closely related in terms of function and regulation. As the membership cutoff decreases, additional genes will be assigned to each cluster group: in many cases, the similarity in the expression of the selected genes will exist over only a subset of the microarray experiments, promoting the identification of conditionally coregulated genes or genes whose products are more peripherally associated with the same cellular processes (see below). The appropriate membership cutoff will vary for each cluster and for the desired results, and selecting meaningful cutoffs can be guided by additional information (see Discussion). For many of the clusters discussed below, the membership cutoffs were empirically chosen to select genes that had coherent gene expression patterns over a given subset of the experiments.

Optimization of fuzzy clustering parameters

The parameters used for the fuzzy clustering were empirically defined for the analysis of yeast genomic expression data. The parameters were optimized to maximally recover clusters identified by hierarchical clustering: these were defined as all hierarchical gene clusters that had a Pearson correlation greater than 0.7 (see Materials and methods); in essence, these clusters served as positive controls. We also assessed the ability of the fuzzy k-means algorithm to identify groups of genes with coherent expression patterns, sets of genes whose products are functionally related, and clusters of known transcription factor targets. A summary of the parameter optimization is discussed below, with additional information available at [14].

Clustering cycles

We found that performing three clustering cycles, with the second and third cycles performed on subsets of the data as described in Materials and methods, maximized the recovery of the clusters identified through hierarchical clustering. Performing three cycles to identify k = 100 centroids recovered 79% of the known clusters in the dataset, compared to the case when the clustering was carried out in one round using identical parameters and seed vectors, for which 64% of the known clusters were identified (see [14]). Performing more than three rounds of clustering did not identify additional known clusters in the dataset, nor did it lead to the increased identification of large clusters of coherently expressed genes (data not shown). We therefore implemented three cycles of clustering, although more sophisticated methods of determining the optimal number of cycles can be envisioned.

Defining k

A significant challenge in partitioning-clustering techniques is defining the number of clusters, k [15]. With standard implementations of the k-means algorithm, underestimating k will result in large clusters of many genes that display divergent gene-expression patterns, while overestimating k will over-fit the data and split groups of similarly expressed genes into multiple, small clusters. Because of the dependence of k-means clustering on k, a number of methods have been developed to estimate this parameter [16,17]. In contrast, fuzzy k-means clustering appears to be less sensitive to over-fitting, because the genes are not forced to belong to only a single cluster. For example, performing the clustering with k = 300 added only approximately 30 unique centroids relative to when the clustering was performed with k = 120, and otherwise-identical parameters (see [14]). The relatively small number of centroids added when k was increased to 300 was largely due to the fact that the program identified many more replicates of centroids, which were consequently removed from the final set. Of the approximately 30 added centroids, most appeared to represent local minima, as they were centroids that were poorly reproduced in bootstrapping experiments (see Materials and methods) and identified few genes that had coherent patterns of expression (data not shown). Nonetheless, the addition of these patterns did not significantly affect the relative memberships of genes to the other centroids (data not shown), indicating that overestimating k did not appreciably affect the clustering results. This presents a significant advantage over standard k-means clustering as it reduces the requirement of accurately estimating k by allowing this parameter to be overestimated.

Initialization

We examined a number of different initialization methods (data not shown) and found that seeding prototype centroids with the eigen vectors identified by PCA performed optimally. Together, the eigen vectors describe the variation in the gene-expression dataset, and therefore seeding the

centroids with these vectors provides a systematic method of sampling the data space. In addition, this protocol produces deterministic clustering results, in contrast to the random initialization method commonly implemented in k-means clustering. One potential drawback of this method is that the number of clusters, *k*, is limited to the number of eigen vectors (which is determined by the number of microarray experiments analyzed). This limitation is alleviated by performing successive cycles of fuzzy k-means clustering on subsets of the data and recalculating the eigen vectors for the respective dataset used in each cycle. In addition, the clustering protocol can incorporate user-defined vectors to seed any number of additional centroids.

Many of the eigen vectors identified by PCA seemed to contain little information about the dataset, as previously noted for this type of analysis [18,19]. Nonetheless, most of the eigen vectors diverged to different gene-expression patterns within 10-15 iterations. The final centroids identified by the fuzzy clustering method showed little dependence on the eigen vectors used to seed the process, as evidenced by bootstrapping analysis. More than 50% of the final centroids were identified in 90% of the bootstrapping trials in which PCA was performed on a random sample of the data, despite the fact the most of the eigen vectors were significantly different in each trial (see Materials and methods). Most of the final centroids bore little similarity to the eigen vectors used to initialize the process, with less than 5% of final centroids similar to any of the eigen vectors with a Pearson correlation greater than 0.7.

Data context

Similarly to standard k-means clustering, the results of the fuzzy k-means method were affected by the data context. This was evident by the fact that the recovery of known clusters was enhanced by performing successive rounds of clustering on data subsets, as described above. In addition, the algorithm performed slightly better on an input dataset that consisted of the subset of yeast genes that showed differential expression patterns, as opposed to the entire gene-expression dataset. As the input dataset for the clustering process, we empirically selected genes whose standard deviation of expression was greater than around 1.4 ($\log_2 0.45$) from each gene's expression mean, amounting to approximately 4,400 out of the approximately 6,200 genes. The algorithm performed equally well on input datasets selected by other criteria of differential expression (data not shown). Performing the clustering on data subsets posed no limitation to the method, because at the end of the procedure all genes in the complete dataset were assigned membership values to the superset of identified centroids.

Fuzzy clustering of yeast genomic expression data

We applied the modified fuzzy k-means algorithm to the analysis of 93 published microarray experiments, each measuring the changes in transcript abundance of the approximately 6,200 predicted yeast genes as cells responded to zinc starvation [4], phosphate limitation [5], DNA-damaging agents [20], and a variety of other stressful environmental conditions [2]. Because the algorithm was not significantly affected by overestimating *k* (described above), we approximated *k* to be roughly double the number of expected clusters in the dataset (defined as the number of clusters identified by hierarchical clustering). Using $k = 120$ and the parameters described in Materials and methods, the algorithm identified 91 unique centroids (Figure 4a; the complete results can be viewed at [14]). More than half of these centroids were correlated more than 0.7 to the cluster means identified by hierarchical clustering of the data, accounting for 87% (46/53) of all of the known clusters in the dataset.

Fuzzy clustering identified previously unrecognized gene clusters

The fuzzy clustering method was also able to identify clusters of genes that were not identified by hierarchical or standard k-means clustering. For example, one centroid (cluster 61, Figure 5a and see [14]) represented genes that were strongly repressed in response to prolonged nitrogen starvation but induced by treatment with the sulfhydryl-rearranging drugs dithiothreitol (DTT) and diamide. Six of the eight characterized genes within the top 20 genes in this cluster encode proteins that are localized to or function in relation to the cell wall, including those involved in bud growth and cell separation (*AXL2*, *CIS3*, *SIM1*), cell-wall proteins induced by antifungal drugs (*SVS1*, *PRY2*, *PRY4*, *CRH1*, *TOS6*), a putative cell-wall sensor (*WSC2*), and a Golgi mannose transporter required for glycosylation of cell-wall proteins (*VRG4*) (see [21] for references). Given the similarity in the expression of these genes, many of the uncharacterized genes may also be involved in processes related to the cell wall. In fact, nearly 70% of all of the genes in this group encode proteins predicted to contain signal peptides (including nine of the twelve characterized genes and five of the eight uncharacterized genes in the group; data not shown [22]), supporting the notion that these proteins are secreted to the cell surface. Extending this group to the top 100 genes in the cluster identified many more similarly expressed genes that are functionally related. In addition to the genes involved in cell wall biosynthesis, which accounted for 30% of the characterized genes selected in this group, an additional 30% of the characterized genes encode proteins involved in protein glycosylation and secretion, while the remaining genes are involved in protein synthesis, cytoskeletal functions, and sterol and lipid biogenesis, all of which can be related to cell-wall and membrane synthesis. Most of the genes that had high memberships in this cluster did not fall into a discrete cluster when the data were analyzed with other clustering algorithms: the top 20 genes associated with this group were distributed into five different clusters when the data were organized by hierarchical clustering and four clusters using k-means clustering (see Materials and methods). This example demonstrates the utility of our

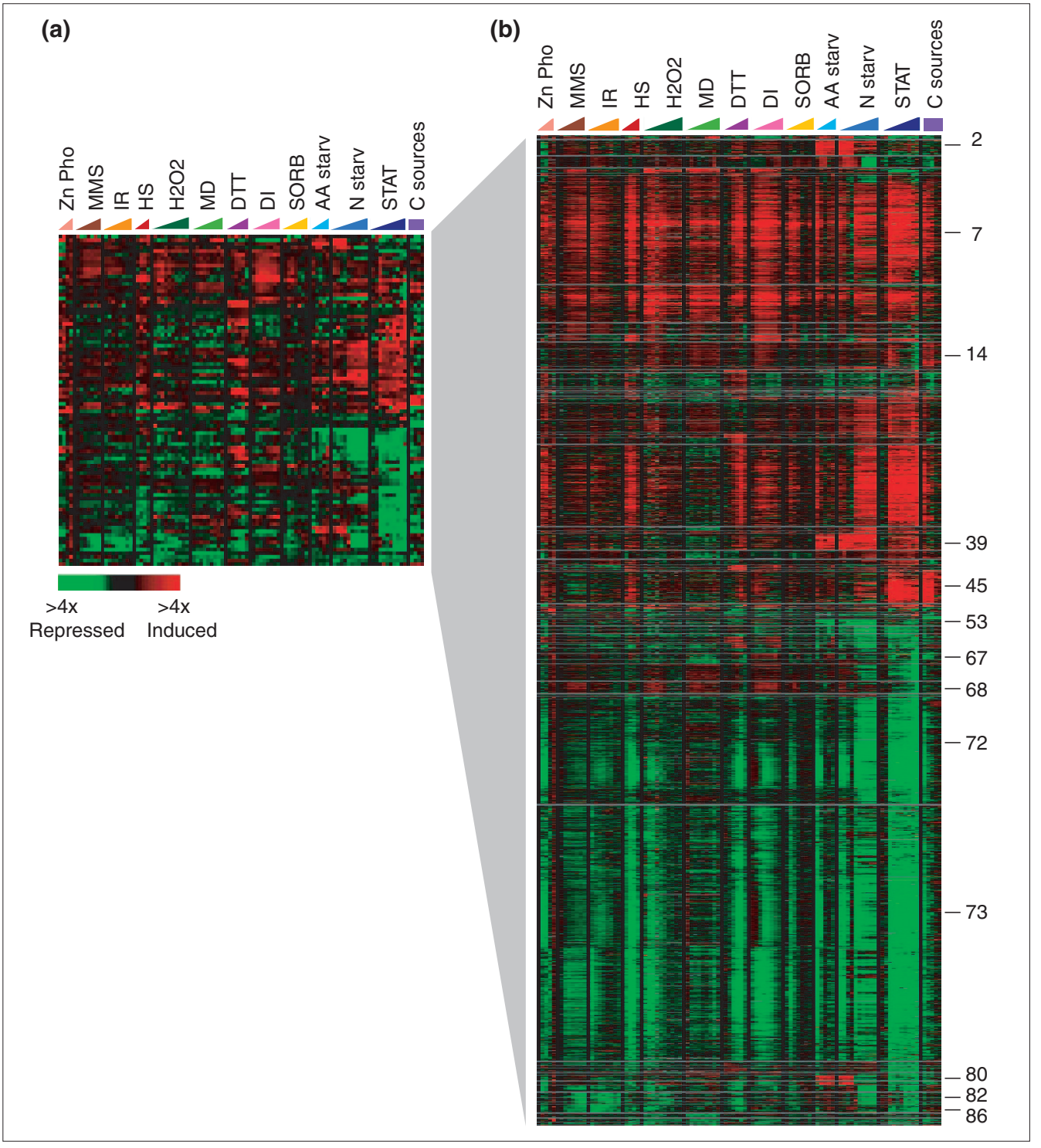


Figure 4
Fuzzy clustering of yeast genomic expression data. **(a)** Each row in this diagram represents one of the 91 centroids identified by fuzzy k-means clustering. The data representation is the same as described in Figure 1. **(b)** Genes were assigned to each of the centroids shown in (a) using a membership cutoff of 0.08, as described in the text. Each cluster of selected genes is separated by a horizontal gray line. Examples of functionally related clusters of genes are indicated by numbers to the right of (b): cluster 2, amino-acid biosynthesis genes; cluster 7, genes induced as part of the environmental stress response; cluster 14, mitochondrial protein synthesis genes; cluster 39, genes involved in nitrogen utilization; cluster 45, oxidative phosphorylation and respiration components; cluster 53, specific amino-acid transporters; cluster 67, glycolysis genes; cluster 68, proteasome components; cluster 72, secretion, protein synthesis, and membrane synthesis genes; cluster 73, genes repressed as part of the environmental stress response; cluster 80, amino-acid biosynthesis genes; cluster 82, G2/M cyclins; cluster 86, histone genes. The complete clustering results can be viewed at [14].

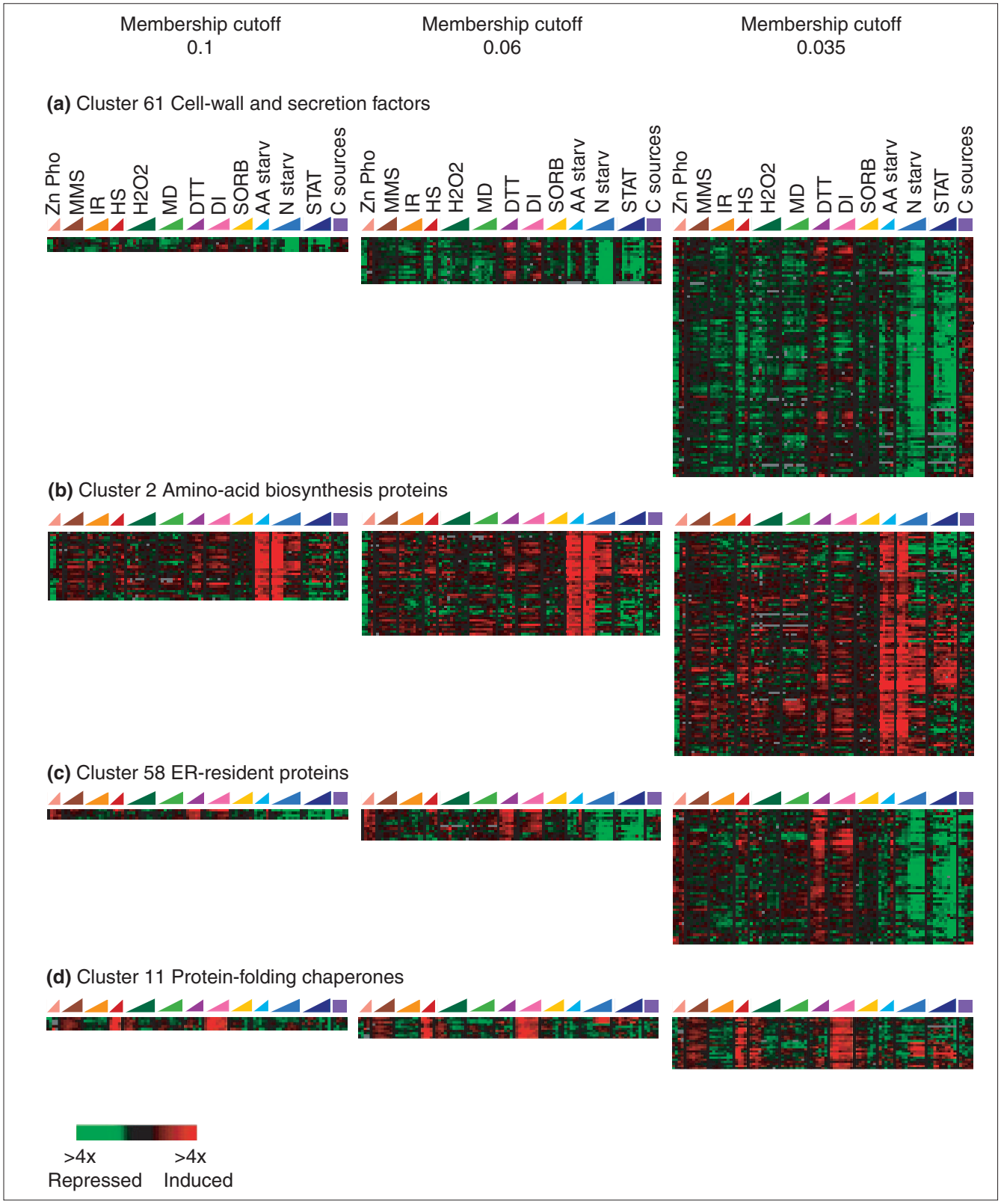


Figure 5
Gene-cluster assignments based on sliding membership cutoffs. Genes that have membership in clusters 61, 2, 58, and 11 greater than the empirically derived membership cutoffs of 0.1 (left), 0.06 (middle), or 0.035 (right) are shown. The data representation is the same as that described in Figure 1. The genes selected in each cluster were hierarchically clustered for display in this figure.

method in identifying previously unrecognized groups of functionally related genes.

In addition to identifying new groups of similarly expressed genes, fuzzy k-means clustering also provided more comprehensive clusters of previously recognized groups of functionally related genes. In many cases, these genes were similarly expressed in only a subset of the experiments, a feature that prevented their association when the data were analyzed with the other clustering methods. An example of this is a centroid that represents genes that were strongly induced by amino-acid starvation (cluster 2 in Figure 5b). Essentially all of the top seven characterized genes associated with this cluster function in methionine biosynthesis and showed similar expression patterns in response to all of the experimental conditions (see [14]). However, as the membership cutoff was decreased to expand the cluster, additional functionally related genes were included, despite the fact that these genes were divergently expressed in response to conditions other than amino-acid limitation. Of the characterized genes within the top 100 genes belonging to this cluster, 64% (42/66) encode proteins that are directly involved in amino-acid biosynthesis, while more than half of the remaining characterized genes are involved in aspects of nitrogen and carbon metabolism that support amino-acid synthesis. Only half of these genes fell into the same cluster when the data were analyzed with hierarchical clustering or k-means clustering (see Materials and methods), while the remaining genes fell into multiple smaller groups in both cases.

Many genes were assigned to multiple clusters

One of the most significant advantages of fuzzy k-means clustering is that genes can belong to more than one group, revealing distinct aspects of their function and regulation. An illustration is provided by the gene *KAR2*, which encodes an HSP70 protein-folding chaperone localized to the endoplasmic reticulum (ER) that is known to respond to defects in ER secretion and to unfolded proteins in this organelle (see [23] for review). Consistent with the known functions of the protein, *KAR2* has significant membership in two clusters. The first (cluster 58 in Figure 5c) includes genes that were induced by the reducing agent DTT, a condition that prevents proper disulfide-bond formation and secretion in the ER [24]. More than 75% (23/31) of the characterized genes within the top 50 genes in this cluster are localized to the ER and participate in various aspects of secretion, including protein folding (*KAR2*, *LHS1*, *FKB2*, *JEM1*), protein disulfide isomerization (*EUG1*, *PDI1*, *ERO1*), protein glycosylation (*GFA1*, *PMT3*, *PMI40*, *SEC59*, *WBP1*, *OST2*), and forward and retrograde trafficking (*ERD2*, *ERP1*, *ERP2*, *SEC24*, *SEC13, RET2, RET3,* and others [21]). Many of the uncharacterized genes in this group are likely to be functionally related to the characterized genes. In addition, *KAR2* also has significant membership in a second cluster (cluster 11 in Figure 5d), which is composed of genes that were induced following heat shock and diamide treatment. Roughly 40% (14/36) of the top characterized genes associated with this group encode protein-folding chaperones localized to different subcellular regions (including those that encode the cytosolic Hsp90 and Hsp70 factors, the mitochondrial Hsp10/Hsp60p and Ssc1p, and the ER- and mitochondrial-associated Ssa1p), and their induction following heat shock and diamide treatment is likely to be in response to widespread protein unfolding inflicted by these conditions. That *KAR2* clusters with both groups of genes reflects the dual role of Kar2p in the response to ER-specific challenges and to conditions that generally destabilize proteins throughout the cell, presumably without affecting other aspects of secretion.

The clustering of *KAR2* with genes in these two clusters not only reflects the functional role of the encoded protein but also corroborates the conditional regulation of *KAR2* expression. In response to defects in ER secretion, *KAR2* is known to be induced by the transcription factor Hac1p as part of the unfolded protein response (UPR) [25-28]. In fact, nearly all of the top 50 genes in cluster 58 were shown by Travers *et al.* [6] to be induced following DTT treatment in a manner dependent on Hac1p and its upstream regulator, Ire1p [6,29,30]. However, unlike most of the genes in cluster 58, *KAR2* is also induced in response to heat shock, along with the other chaperone genes in cluster 11, by the transcription factor Hsf1p [31]. Consistently, most of the top genes in this group, including *KAR2*, contain multiple Hsf1p-binding sites in their promoters. The clustering of *KAR2* with both clusters of genes therefore reflects the known induction of the gene by Hac1p as part of the UPR but by Hsf1p following heat shock.

Many additional yeast genes have significant membership in more than one of the fuzzy clusters. When the genes were assigned to all clusters with an empirically defined membership cutoff of 0.06, more than a third of the assigned genes were placed in more than one group (Table 1); at a slightly lower cutoff of 0.04, almost two-thirds of all of the assigned genes were placed into multiple clusters. As with *KAR2*, the fuzzy assignment of many of these genes was consistent with the known roles of the encoded proteins. Genes involved in histidine biosynthesis (for example *HIS4* and *HIS5*) clustered with other genes involved in amino-acid synthesis (cluster 2 and cluster 80) but also with genes required for adenine biogenesis (cluster 1), in agreement with the roles of these gene products in both histidine and purine metabolism [32,33]. Genes that encode ER vesicle coat proteins (*SEC13*, *SEC21*, *SEC24*, *COP1*, *RET2*, *RET3* and others) were induced with other ER-specific genes in response to the UPR, as discussed above for cluster 58, but were also strongly repressed following long-term nitrogen and carbon starvation, along with hundreds of other genes that function in diverse aspects of secretion and protein synthesis (cluster 72). The repression of these genes coincided with the cellular growth arrest resulting from the starvation conditions and was likely to be

Table 1

Fuzzy assignment of genes to clusters

Membership cutoff*	Number of genes assigned†	Number of genes assigned to >1 cluster‡	Percent assigned genes in >1 cluster§
0.10	1,341 (22%)	230 (4%)	17%
0.08	1,843 (30%)	334 (5%)	18%
0.06	2,631 (43%)	913 (15%)	35%
0.04	4,233 (69%)	2,719 (44%)	64%

*The membership cutoff listed was used to assign genes to all of the clusters. †Number and fraction of the total number of genes that were assigned to any of the 91 clusters described in the text. ‡Number and fraction of the total number of genes that were assigned to more than one cluster. §Fraction of the placed genes that were assigned to more than one group.

triggered by the decreased demand for protein and membrane synthesis in nondividing cells. That these genes belong to multiple, distinct clusters reflects the condition-specific roles of the encoded proteins and suggests that their conditional expression with these alternative groups of genes is triggered by different cellular signals.

Fuzzy clusters represent gene targets of yeast transcription factors

Many of the genes that clustered together by fuzzy k-means clustering are likely to be coregulated at the level of transcription in response to certain environmental conditions. We explored this possibility by characterizing the enrichment of each fuzzy cluster for genes that contained known transcription factor binding sites within 800 base-pairs (bp) upstream of their open reading frames (ORFs). Genes were assigned to each cluster using an empirically chosen membership cutoff of 0.06 or 0.08, and the probability of observing the number of genes in each cluster that contained one or more copies of each transcription factor binding site was calculated, based on the hypergeometric distribution (see Materials and methods). Roughly 25% of the identified fuzzy clusters were statistically enriched (with $P < 2 \times 10^{-4}$) for genes that contain copies of at least one of 43 different promoter elements, and more than half of these clusters were enriched for multiple sites (the complete results are available at [14]). In many cases, the presence of the promoter elements was consistent with the known regulation of the genes' expression (Figure 6). For example, cluster 2 was enriched for genes that contain binding sites for Gcn4p, Cbf1p, and Met31/32p. Around 75% of these genes are known to be induced by Gcn4p in response to amino-acid limitation and contain the Gcn4p promoter element [34,35]. However, those that are specifically involved in methionine synthesis also contain the recognition sequences for Met31/32p and/or Cbf1p, factors that cooperatively regulate gene expression according to the demand for the products of this pathway [36-39].

At a membership cutoff of 0.08, cluster 45 consisted of many genes involved in the tricarboxylic acid cycle and oxidative phosphorylation, and this group was enriched for the binding site of the Hap2/3/4p complex that is known to regulate the genes' expression. At a slightly lower cutoff of 0.06, additional genes involved in respiration and utilization of alternative carbon sources were assigned to the cluster, making the enrichment of the promoter element recognized by the catabolite repressor Mig1p statistically significant [40-42]. At both of these membership cutoffs, this cluster was also highly enriched for the sequence recognized by the stress-responsive factors Msn2p and Msn4p. These factors recognize a sequence that is very similar to the Mig1p-binding site, and it is possible that the enriched sequence actually represents derivative Mig1p elements. However, a specific role for Msn2p in the response to glucose starvation has recently been identified [43], raising the possibility that the factor is directly involved in regulating these genes. Another cluster (cluster 73) consists largely of genes that were sharply repressed in response to environmental stresses [2,44]. The ribosomal protein genes in this group are regulated by the factor Rap1p and contain multiple copies of its binding site within their promoters [45,46], while other genes in this group contain two putative regulatory sequences that have been previously identified as enriched in the promoters of many of these genes [2,47-50]. In each of these cases, the comprehensive clusters identified by the fuzzy k-means analysis included additional genes that were not previously known to be targets of these factors. Some of the binding-site occurrences shown in Figure 6 are likely to have occurred by chance (especially for sequences that are common in the genome, such as the Hap2/3/4p-, Mig1p-, and Msn2/Msn4p-binding sites). However, the similarity between the expression patterns of these genes and those of the known transcription factor targets, along with the functional correlation of the gene products and the presence of the respective binding sites in the genes' promoters, strongly suggest that many of these genes are legitimate targets of these regulators.

The overlapping clusters identified by fuzzy k-means clustering presented more complete groups of transcription factor targets compared to other clustering methods, enhancing the identification of promoter elements enriched in the clusters and implicating details of the conditional regulation of gene expression. An example can be seen in a set of around 15 genes that are induced by Yap1p in response to oxidative stress, but by the general stress factors Msn2p and/or Msn4p (Msn2/Msn4p) in response to other stressful conditions [2]. These genes belonged significantly to three different clusters (Figure 6). One cluster (cluster 7) consisted of around 90 known Msn2/Msn4p targets and was enriched for genes whose promoters contain the known Msn2/Msn4p-binding site as well as other C-rich sequences that are similar to, but distinct from, the Msn2/Msn4p site (see below). The second cluster that these

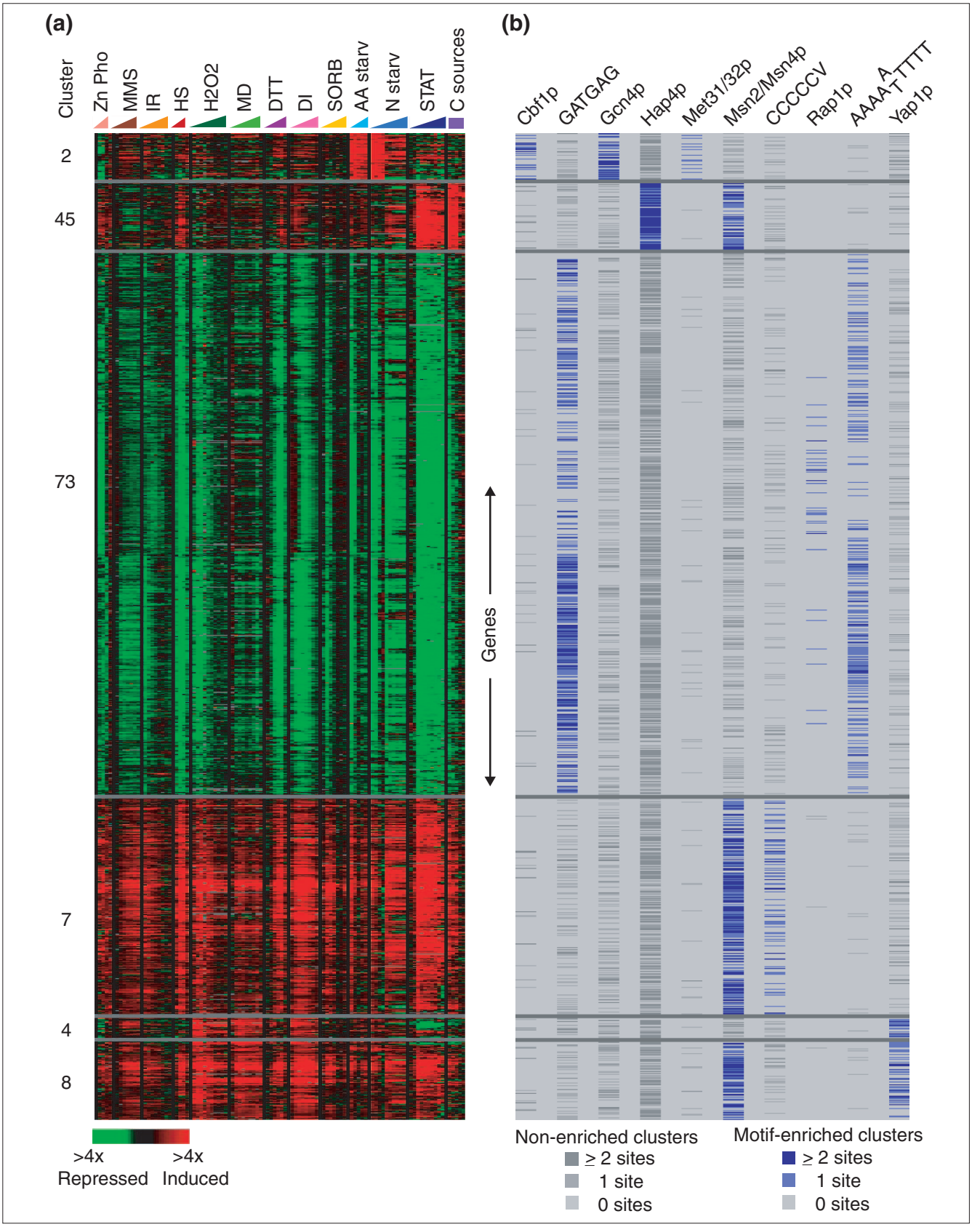


Figure 6 *(see the legend on the next page)*

genes belonged to (cluster 4) comprised known Yap1p targets, most of which contain the Yap1p-binding site within their promoters. The third cluster (cluster 8) specifically represented the subgroup of genes that are conditionally regulated by Yap1p or Msn2/Msn4p, and this group was enriched for both the Yap1p element and the Msn2/Msn4p-binding site (but not other C-rich sequences). At a lower membership cutoff, additional genes were assigned to cluster 8 that showed similar expression patterns and contain both the Yap1p and Msn2/Msn4p promoter elements, suggesting that these genes may also be conditionally regulated by the factors. In contrast to these results, when the data were analyzed by k-means clustering, these genes could only be assigned to a cluster of Yap1p targets or a cluster of Msn2/Msn4p targets, and therefore no group of genes that was statistically enriched for both of these promoter elements could be identified (data not shown).

The majority of the clusters identified by fuzzy k-means clustering were not statistically enriched for known transcription factor binding sites. To identify novel enriched promoter sequences, we calculated the hypergeometric distribution of all possible 6-mer sequences in the promoters of the genes clustered by fuzzy k-means clustering. Almost all the statistically significant 6-mers represented known transcription factor binding sites, with the exception of a group of C-rich sequences with high statistical enrichment in the promoters of the Msn2/Msn4p targets in cluster 7 (Figure 6). We therefore focused our attention on the newly identified group of cell-wall genes, defined as the top 20 genes in cluster 61. Although none of the 6-mers met the significance cutoff for this cluster ($P = 10^{-6}$), the most significant sequence (CGCGAA, $P = 10^{-5}$) was identical to the core binding site of SBF, a transcription factor complex that regulates cell-cycle-dependent gene expression at the G1 to S transition [51-53]. In fact, more than two-thirds of these genes were identified by Iyer *et al*. [54] as part of a larger set of around 180 genes whose flanking regions were physically bound by the SBF complex. However, this set of genes was not coordinately expressed during cell-cycle progression [54,55], suggesting that these cell-wall genes may be regulated by a distinct mechanism in response to environmental conditions.

To try and identify novel sequences conserved in these promoters, we used the motif-finding algorithm MEME [56], initializing the EM algorithm with the most significantly enriched 6mer for the cluster. Two sequences were repeatedly identified using a variety of parameters (Figure 7): one motif was very similar to, but extended from, the known SBF-binding site and was present in 80% of the gene promoters, often in multiple copies. This represents a significant enrichment over the entire set of yeast gene promoters, of which approximately 30% contain the site. Nearly 90% of the promoter regions that contain this motif were shown to be physically bound by the SBF complex [54], consistent with the idea that SBF binds this sequence. These details raise the possibility that SBF coordinates the expression of this set of cell-wall genes in response to conditions other than cell-cycle progression [54], perhaps in a manner that is specifically dependent on the variant site identified here. MEME also identified another sequence that has not previously been implicated in gene-expression regulation (Figure 7b). More than 55% of the top 20 genes in cluster 61 harbor the motif in their promoters, in contrast to only 16% of all yeast promoters. The precise roles of these sequences in mediating gene expression will require further experiments; however our ability to identify novel sequences conserved in these promoters highlights the potential for discovering additional regulatory elements by fuzzy k-means clustering.

Fuzzy clustering uncovered correlations between the environmental conditions

As well as understanding the similarities between genes based on their expression patterns, it is also enlightening to correlate the experimental conditions in terms of their effects on gene expression. These correlations can implicate common features of the environments while pointing to the regulatory systems that are activated in each situation. A convenient feature of the fuzzy-clustering output is that the experiments can be hierarchically clustered, on the basis of the expression patterns of each selected subset of genes, to reveal similarities in the effects of the conditions. Performing the clustering based on subsets of genes presents subtle correlations between the experiments that cannot be realized when the clustering is based on all the genes in the

Figure 6 *(see the figure on the previous page)*
Fuzzy clusters are enriched for genes that contain known transcription factor promoter elements. The enrichment of each cluster for genes that contain known transcription factor binding sites in their promoters was measured on the basis of the hypergeometric distribution, as described in Materials and methods. **(a)** Gene-expression data (as described in Figure 1) for genes that were assigned to cluster 2 (amino-acid biosynthesis genes), cluster 45 (respiration genes), cluster 73 (genes repressed as part of the environmental stress response (ESR)), cluster 7 (genes induced as part of the ESR), cluster 4 (oxidative stress defense genes), and cluster 8 (genes conditionally regulated by Yap1p or Msn2/Msn4p). The genes were assigned to each cluster with a membership cutoff of 0.08, with the exception of cluster 2 for which a cutoff of 0.06 was used. The hypergeometric distribution was used to measure the statistical enrichment of promoters containing the binding sites of Cbf1p (TGACGTG), the ESR motif GATGAG, the binding site of Hap2/3/4p (CCAAT), Met31/32p (AAACTGTG), Msn2/Msn4p (CCCCT), a C-rich element identified in cluster 7 (CCCCCV where V is any nucleotide but T), Rap1p (ACACCCAYACAY where Y is C or T), the ESR motif AAAAWTTTT (where W is A or T), and Yap1p (TTAGTMA where M is C or A). **(b)** For each gene displayed in (a), the copy number of the denoted transcription factor binding sites in the gene's promoter is indicated by a colored box. The copy number is indicated with a blue box only if the cluster to which the gene belonged was statistically enriched ($P < 2 \times 10^{-4}$) for the indicated binding site, whereas the copy number is indicated with a dark-gray box if the cluster to which the gene belonged was not statistically enriched for the site. The complete results are available at [14].

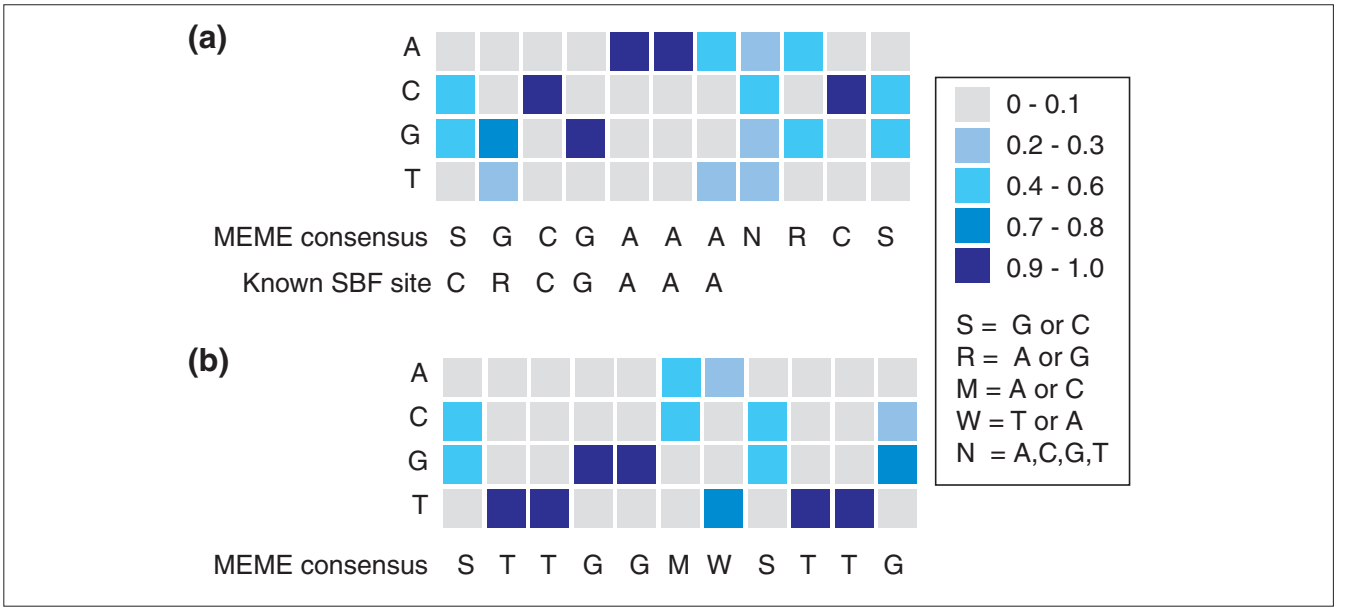


Figure 7
Motifs conserved in the promoters of the cell-wall genes. Two motifs were identified in the promoters of the cell-wall genes selected in cluster 61, using MEME [56]. The position weight matrices are represented by colored boxes, where each box indicates the frequency of the denoted nucleotide at that position in the matrix, according to the color key shown. The position weight matrices and consensus sites are shown for **(a)** the SBF-like sequence and the known SBF consensus site, and **(b)** the novel sequence identified in these promoters.

dataset. For example, when the 93 microarray experiments analyzed in this study were hierarchically clustered on the basis of the expression patterns of all the genes in the dataset, the experiments largely clustered according to the individual time courses (Figure 8a). This reveals that the overall genomic expression program triggered by each environment was unique to each set of conditions.

However, when the microarray experiments were clustered on the basis of subsets of genes identified by fuzzy k-means clustering, more detailed correlations emerged, indicating more information about the effects of each environment. An example is the sulfhydryl-oxidizing drug diamide, which affects many aspects of cell biochemistry. When the experiment clustering was performed on the basis of genes encoding protein-folding chaperones (the top 10 genes in cluster 11), a striking similarity between the effects of diamide and heat shock was observed (Figure 8b). In contrast, when the microarray clustering was performed on the basis of genes involved in oxidative stress defense, (the top 24 genes belonging to cluster 4), diamide was most similar to hydrogen peroxide and menadione, which inflict oxidative damage by generating reactive oxygen species (Figure 8c). In terms of the genes induced in the UPR (identified as the top 30 genes or so in cluster 58), the effects of diamide were most similar to those triggered by the reducing agent DTT (Figure 8d). That the effects of diamide were similar to those of different environmental conditions depending on the genes analyzed reflects the diverse effects of this drug on the cell (Figure 8f). By crosslinking protein sulfhydryl groups, diamide is thought to disrupt protein structure and trigger oxidative stress [57,58], both of which are likely to perturb normal ER functions [59].

The similarities between other environmental conditions are less well understood. For example, limitation of the essential nutrient zinc triggered diverse gene-expression changes in the cell [4]. Although the overall genomic expression

Figure 8 *(see the figure on the next page)*
Differential hierarchical clustering of the conditions based on different fuzzy gene clusters. **(a)** The dendrogram generated by hierarchically clustering the experimental conditions based on all of the genes in the dataset is shown. **(b-e)** Portions of the dendrograms generated by hierarchically clustering the experiments based on (b) protein-folding chaperones and other genes assigned to cluster 11 (top 10 genes), (c) oxidative stress genes assigned to cluster 4 (top 24 genes), (d) UPR genes assigned to cluster 58 (top 33 genes), and (e) genes involved in respiration and carbon metabolism assigned to cluster 36 (top 96 genes). **(f)** A summary of the clustering results (discussed in detail in the text), where each arrow indicates the induced expression of the respective gene set in response to the conditions indicated. The known regulators of genes represented in each cluster are shown. The complete results are available at [14].

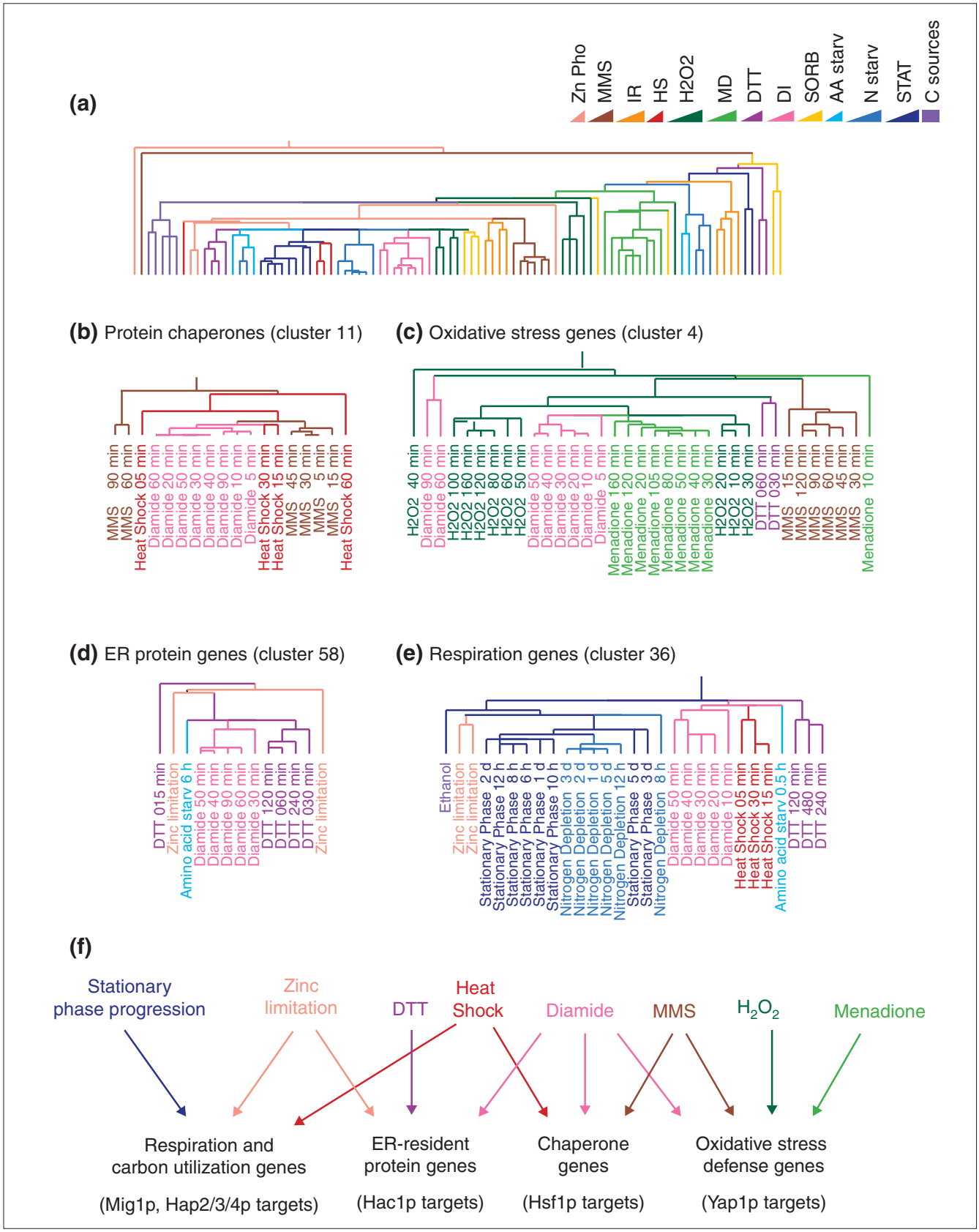


Figure 8 *(see the legend on the previous page)*

program triggered by this condition was distinct, when the experiment clustering was based on genes involved in the use of alternative carbon sources and respiration (cluster 36 and cluster 45, respectively), a significant similarity between the effects of zinc limitation and conditions that involve glucose starvation emerged (Figure 8e). Like carbon starvation, zinc limitation triggered the increased expression of these genes, even though the cells were not limited for glucose (T. Lyons, personal communication). This result suggests that zinc-limited cells may have a defect in glucose metabolism, leading to the induced expression of the respiration and carbon-utilization genes. While a link between zinc limitation and sugar metabolism has been established in mammals [60], the molecular basis of this correlation is not known. In contrast to this relationship, when the experiment clustering was performed on genes encoding ER-resident proteins, the effects of zinc starvation were most similar to those inflicted by DTT and diamide (Figure 8d), suggesting that zinc limitation may initiate the UPR. Zinc starvation is not known to induce this program in yeast (C. Patil, personal communication). However, a potential connection between zinc and the UPR is the protein calreticulin, an ER protein-folding chaperone that acts on glycosylated proteins [61]. That mammalian calreticulin is a zinc-dependent protein [62] raises the possibility that zinc limitation prevents the proper activity of this protein in yeast, leading to unfolded ER proteins and triggering the subtle increase in expression of genes that participate in the UPR. The correlations between the gene-expression changes triggered by these conditions suggest hypotheses about the effects of each condition that warrant future experiments.

Discussion

To respond to diverse and frequently changing environmental conditions, yeast cells must precisely mediate the synthesis and function of the proteins in the cell. This is controlled in part by the overall genomic expression program that results from the combined action of different regulatory factors, each of which responds to specific extra- and intracellular signals. Many of these regulators act under specific conditions, and together they govern the expression of overlapping sets of genes. Individual genes, in turn, are regulated by multiple, condition-specific systems that result in each gene being coexpressed with different groups of genes under different situations.

Although examples of this type of regulation have been observed on an individual gene basis, our results suggest that the condition-specific regulation of overlapping sets of yeast genes is a prevalent theme in the regulation of yeast gene expression. A large fraction of yeast genes is expressed in patterns that are similar to different groups of genes in response to different subsets of the experiments (Table 1). Furthermore, a substantial number of these genes contain multiple transcription factor binding sites in their promoters (Figure 6, and see [14]), consistent with the idea that they are conditionally regulated by multiple, independent regulatory systems. The condition-specific regulation of gene expression has also been implicated in higher organisms [63,64] and probably has a significant role in regulating genomic expression. This is in contrast to the regulatory logic of prokaryotes, in which the expression of defined sets of genes in operons is a predominant feature of regulation. Thus, the conditional regulation of overlapping groups of genes may represent a regulatory theme that is particularly important in eukaryotes.

The prevalence of conditional gene coexpression poses a challenge for the analysis of gene-expression data, because many genes will have expression patterns that are similar to multiple, distinct gene groups. Fuzzy k-means clustering is well suited to identifying conditionally coexpressed genes for a number of reasons. First and foremost, the method can present overlapping clusters, revealing distinct features of each gene's function and regulation. The resulting implications can be used to assign refined hypothetical functions to uncharacterized gene products on the basis of the known functions encoded by the genes in each cluster. In addition, this information can suggest additional cellular roles of well studied proteins (see [14]). The overlapping clusters identified by fuzzy k-means clustering also present more comprehensive groups of conditionally coregulated genes. This is especially important for the successful identification of regulatory motifs common to the promoters of similarly expressed genes, because motif-finding algorithms are often hindered by small sample sets. More than two-thirds of the gene clusters we identified are not enriched for known regulatory elements, highlighting the potential for discovering novel sequences involved in gene-expression regulation. We expect that fuzzy k-means clustering will advance that discovery, as illustrated by our ability to identify new sequences conserved in the promoters of clustered genes.

Another benefit of the fuzzy k-means algorithm is that it identifies continuous clusters of genes. This allows each cluster to be expanded or collapsed to view genes of varying similarity in expression. While the genes of highest membership in a given cluster are often tightly correlated in terms of biochemical function and regulation, expanding the cluster can identify genes that are similarly expressed in only subsets of the experimental conditions. The resulting gene relationships can suggest details about the cellular roles served by the encoded gene products and the regulatory systems that govern the genes' expression in response to the relevant conditions. Thus, the results of fuzzy k-means clustering are naturally suited for biologists to use in an intuitive and physiologically meaningful way.

The unique features of fuzzy k-means clustering have allowed us to uncover complex similarities in yeast gene-expression patterns, identify putative transcription factor

binding sites present in the genes' promoters, and elucidate the environmental conditions that trigger changes in gene expression. Integrating these details can indicate the cellular signals and regulatory systems that govern the expression of specific sets of genes in yeast (Figure 9). For example, the fuzzy clustering of genes involved in methionine biosynthesis with other amino-acid biosynthetic genes and with genes involved in nitrogen utilization lead to the identification of multiple transcription factor binding sites in the genes' promoters. Together, these details reflect the alternative regulatory systems that are known to govern the expression of the methionine biosynthesis genes. Although they are induced by one regulatory system (Cbf1p-Met31/32p) according to the demand for the pathway's products, they are induced by an alternative system (Gcn4p) in response to a general signal of amino-acid starvation [34,35,65], and they are probably also regulated by a third mechanism (GATA factors) in response to the available nitrogen source. Combining this information with similar indications for other sets of genes gives a summary of the details discussed in this study and suggests a model for the organization of the regulatory system that controls gene expression in yeast (Figure 9). The overlapping nature of the sets of coregulated genes supports the ability of the cell to customize the emergent genomic expression program to the particular needs of the cell, while minimizing the number of regulators required to produce each genomic expression program.

The fuzzy k-means algorithm used here was chosen for its conceptual and algorithmic simplicity. There are many alternative algorithms that might accomplish the same ends. For example, Ihmels *et al.* [11] have applied a heuristic algorithm to the analysis of yeast gene-expression data to identify overlapping sets of genes whose expression is similar to known gene-expression patterns. This method produced interesting results and identified genes that were similarly expressed to known transcription factor targets. A key difference between these algorithms is that fuzzy k-means clustering requires no *a priori* information about the dataset. Thus, each method may be suitable for a different biological question, namely identifying genes whose expression is similar to known or expected gene expression patterns versus an unbiased, *de novo* exploration of the gene-expression dataset.

Despite the advantages of fuzzy k-means clustering discussed above, the method also has a number of limitations. Most notably, the assignment of genes to the clusters requires a user-defined membership cutoff. While this allows complete flexibility in data exploration, selecting meaningful cutoffs is a challenge. Choice of cutoff can be guided by a number of criteria, including the coherence of the selected gene-expression patterns, the functional relationships of the characterized genes selected, or the statistical enrichment of sequences in the selected genes' promoters. We have attempted to alleviate the challenge of selecting cutoffs by providing visualization software specifically designed for the fuzzy clustering results, allowing the gene expression data to be inspected directly and dynamically.

Although the fuzzy k-means clustering method successfully identified nearly 90% of the known clusters in the dataset, it routinely failed to identify a small number of groups that were identified by hierarchical clustering. The inability of the method to find the expression patterns representing these groups seemed to be dependent on the overall properties of the dataset, rather than the absence of an appropriate eigen vector used to initiate the process, as the program was unable to identify these patterns even when the process was initiated by seeding the centroids with the unidentified patterns (data not shown). We have accounted for this limitation by allowing any number of expression patterns to be added to the final list of identified cluster centroids, thereby revealing genes that are similarly expressed to the pattern in question.

Despite these limitations, the unique advantages of fuzzy k-means clustering make the technique a valuable tool for gene-expression analysis. We believe that fuzzy k-means clustering will be a useful complement to other computational methods commonly used to analyze gene-expression data. Whereas algorithms that present discrete gene clusters provide a straightforward method of initial data exploration, the flexibility of fuzzy k-means clustering can be used to reveal more complex correlations between gene-expression patterns, promoting refined hypotheses of the role and regulation of gene-expression changes.

Materials and methods

Software and supplementary information

The clustering software FuzzyK and the visualization program FuzzyExplorer are available from [14], along with the complete clustering results and additional information.

Dataset

Published genomic expression data of wild-type *S. cerevisiae* responding to zinc starvation [4], phosphate limitation [5], DNA-damaging agents [20], and a variety of stressful environmental changes [2] were combined into a dataset of 6,153 genes and 93 microarray experiments (dataset A). These data were chosen because the experiments were performed using the same experimental and microarray methods [55]. The data were downloaded from the Stanford Microarray Database and were otherwise unprocessed before clustering, with the exception of the heat shock, DTT, and carbon-source experiments, which were transformed as previously described [2]. The complete dataset organized by hierarchical clustering can be downloaded from [14]. A subset of this data was used in the fuzzy k-means clustering and consisted of 4,373 genes whose standard deviation in expression was $\log_2(0.45)$ from each vector mean (dataset B), identified using the program Cluster [66].

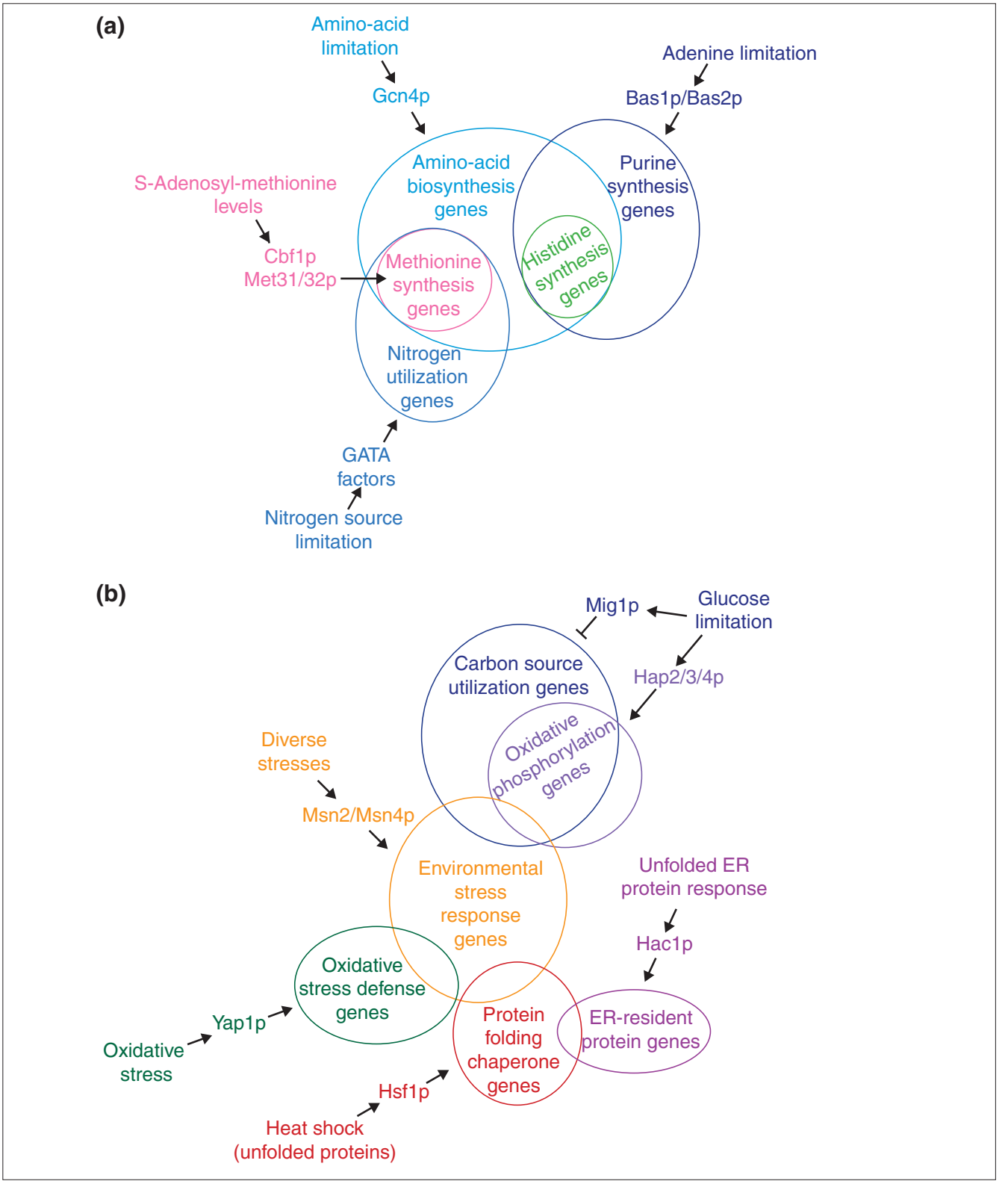


Figure 9
Integration of gene expression, regulatory sequences, and environmental responsiveness. Schematics illustrating the hypotheses presented in this paper regarding the regulation of expression of genes that respond to **(a)** amino-acid and nitrogen source limitation and **(b)** different environmental stresses. Each circle in the Venn diagram represents a cluster of genes that is enriched for the known binding site of the indicated transcription factor or is known to be regulated by the indicated factor in response to the denoted conditions.

Similarity metric

For all methods discussed, the weighted, uncentered Pearson correlation was used as the similarity metric (referred to simply as the Pearson correlation) [1]. Where noted, the Pearson distance was used, equal to 1 – correlation. The array weights used in the calculation were generated as previously described, using a Pearson correlation cutoff of 0.8 and an exponent of 1 (see [67] for details).

Hierarchical clustering

Average linkage hierarchical clustering of the data was carried out using the program Cluster as previously described, using the weighted, uncentered Pearson correlation as the similarity metric [1]. Dataset A and dataset B were hierarchically clustered with identical parameters, using array weights calculated based on a correlation cutoff of 0.8 and an exponent of 1 [67]. Clustering of the microarray experiments was carried out similarly, using gene weights calculated with a correlation cutoff of 0.7 and an exponent of 1.

To represent all the significant gene clusters identified by hierarchical clustering, the dendrogram generated for dataset B by the Cluster program was parsed, and the average expression patterns of clusters of more than three genes with an average Pearson correlation >0.7 were calculated. Through this method, 38 hierarchical cluster means were identified. The parsing process was repeated to calculate the average expression patterns of clusters of more than three genes with average correlations >0.8 and >0.9. Cluster means not already represented in the initial group of 38 clusters were added to the group, resulting in a total of 53 hierarchical cluster means identified for the dataset. The centroids identified through fuzzy k-means clustering were considered similar to the hierarchical cluster means if the Pearson correlation between the vectors was >0.7.

Fuzzy k-means clustering

We implemented the modified fuzzy k-means method in the C++ program FuzzyK, available at [14].

The fuzzy k-means algorithm [12] is based on the minimization of the objective function shown below, for a given fuzzy partition of the data, F, and a set of K cluster centroids, V

$$J(F,V) = \sum_{i=1}^{N}\sum_{j=1}^{K} m_{XiVj}^{2}\, d_{XiVj}^{2}$$

where X_i is the expression pattern of the ith gene in the dataset, V_j is the centroid of the jth cluster, d is the Pearson distance between X_i and V_j, m_{XiVj} is the membership of X_i in cluster V_j, N is the number of genes in the dataset, and K is the total number of clusters.

We implemented the algorithm to perform three successive cycles of fuzzy k-means clustering. The first cycle of clustering was initialized by performing PCA on dataset B using the GNU Scientific Library SVD function. Of the top $k/3$ eigen vectors, those to which no gene had a maximal Pearson correlation were eliminated (for k = 120, only one eigen vector was eliminated in each cycle). The remaining eigen vectors were used as prototype centroids for that clustering cycle. Subsequent cycles of clustering were initialized similarly, except that PCA was performed on the respective data subset used in that clustering cycle.

During the centroid refinement in each clustering cycle, new centroids were calculated on the basis of the weighted mean of all the gene-expression patterns in the dataset according to

$$V_j' = \frac{\sum_{i=1}^{N} m_{XiVj}^{2}\, W_{Xi}\, X_i}{\sum_{i=1}^{N} m_{XiVj}^{2}\, W_{Xi}}$$

where each gene's membership m (a continuous variable from 0 to 1) was defined as

$$m_{XiVj} = \frac{\frac{1}{d_{XiVj}^{2}}}{\sum_{j=1}^{K} \frac{1}{d_{XiVj}^{2}}}$$

and w was the gene weight: in the first clustering cycle, the gene weights used were those defined by the program Cluster, using a Pearson correlation cutoff of 0.7 and an exponent of 1 [67] and in subsequent cycles the gene weight was empirically defined as

$$W_{Xi} = \left(\sum_{n=1}^{N} \frac{c_{Xi,Xn} - x}{1 - x}\right)^{2}$$

where $d_{Xi,\,Vj}$ is the Pearson distance between gene X_i and vector V_j, $c_{Xi,\,Xn}$ is the Pearson correlation between genes X_i and X_n, and x is the correlation cutoff, in this case 0.6. This weighting scheme served to overweight genes that were correlated to other genes in the dataset.

In each clustering cycle, the centroids were iteratively refined until the average change in gene memberships between iterations was <0.001 (approximately 40-60 total iterations in each clustering cycle). While around 85% of the centroids stabilized within approximately 15 iterations, some of the centroids required more than 40 iterations before stabilizing.

After each clustering cycle, the centroids were combined with those identified in previous cycles, and replicate centroids were averaged: each centroid was compared to all other centroids in the set, and centroid pairs correlated >0.9 were replaced by the average of the two vectors. The new

vector was compared to the remaining centroids in the set and averaged with those to which it was correlated >0.9. This process continued until each centroid (or the vector that replaced it) was compared to all other existing centroids in the set.

Following the first and second clustering cycles, data subsets were selected to apply to subsequent rounds of clustering. Genes that were correlated to any existing centroid with a Pearson correlation >0.7 were removed from the dataset, and array and gene weights were recalculated on the data subset as described above. The new data subset was applied to a subsequent cycle of clustering, performed as described above.

Final gene-cluster assignments

The centroids identified through three rounds of fuzzy clustering were combined into one set and replicate centroids were averaged, as described above. Each gene in dataset A was assigned a membership score to each of the unique centroids. For display in the figures, the final list of centroids was ordered by hierarchical clustering. Genes were selected in each cluster if their membership score was greater than the empirically determined membership cutoff applied to each cluster. For display in Figure 5, the genes selected in each cluster were subsequently organized by hierarchical clustering.

k-means clustering

For an optimal comparison of the results of k-means and fuzzy k-means clustering, we performed the k-means clustering identically to the fuzzy k-means protocol, except that during the clustering iterations each gene contributed only to the cluster to which it was most similar (with a membership of 1.0). Three rounds of hard k-means clustering were performed with $k = 120$, and each cycle was initiated by seeding $k/3$ centroids with the most informative $k/3$ eigen vectors identified by PCA, as described above. The process for merging centroids, selecting the data subsets for subsequent clustering rounds, and gene and array weighting were carried out identically as described for fuzzy k-means clustering. After identification of the final set of centroids, each gene was assigned only to the centroid to which it was most similar.

Bootstrapping

To estimate the dependence of the procedure on the initial dataset, a bootstrapping method was applied in which the fuzzy k-means protocol was repeated 100 times, each time on 4,373 genes chosen randomly from dataset B, with $k = 102$. The occurrence of each centroid in the bootstrap trials was determined by summing the number of trials that contained a centroid that was correlated >0.7 to the centroid in question. By this criterion, roughly 50% of the centroids were identified in 90% of the trials, while approximately 25% of the centroids were identified in all of the trials.

To estimate the dependence of the procedure on the eigen vectors used to seed the clusters, a similar bootstrapping procedure was carried out in which PCA was performed on 4,373 genes chosen randomly from dataset B but the cluster refinement was done using all genes in dataset B. The frequency of each centroid was scored as described above. The results were similar to the previous bootstrapping experiment, with around 50% of the centroids present in 90% of the bootstrapping trials, and around 25% of the centroids identified in all the trials.

Promoter analysis

Genes were assigned to all the 91 identified centroids on the basis of a membership cutoff of 0.06 or 0.08. The statistical enrichment of each cluster for genes that contained known transcription factor binding sites or different 6-mer sequences within 800 bp upstream of the ORF was assessed, according to the hypergeometric distribution. The probability of observing at least q genes that contained one or more copies of a given sequence out of l genes in a fuzzy cluster was calculated as

$$\sum_{i=q}^{l} \frac{\binom{M}{i}\binom{N-M}{l-i}}{\binom{N}{l}}$$

where M is the number of genes in the genome that contain the motif and N is the total number of genes in the genome. Forty-three transcription factor binding sites were compiled from the literature (see [14] for a complete list of sequences). The enrichment of each sequence was considered significant if the P value was <0.01 divided by the number of elements searched, or 2×10^{-4} for the 43 transcription factor binding sites and 2×10^{-6} for the 4,096 different 6-mers.

The program MEME [56] was seeded with the most significant 6-mer (CGCGAA) enriched in the promoters of the genes selected for cluster 61, and the program was run with a variety of parameters (the parameters and MEME output can be found at [14]). Genes whose promoters contained significant matches to the identified matrices were identified using Patser on the RSA tools website [68,69].

Acknowledgments

We thank A. Moses and E. Kelley for helpful suggestions and programming assistance, A. Alizadeh, J. Bolderick, N. Ogawa, C. Patil, C. Rees, P. Spellman, and M. Kamvysselis for helpful discussions, and A. Moses, D. Chiang, and J. Fay for critical reading of the manuscript. A.P.G. is supported by an NSF postdoctoral fellowship in biological informatics, and M.B.E. is a Pew Scholar in the Biomedical Sciences. This work was conducted under the US Department of Energy contract No. ED-AC03-76SF00098.

References

1. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.
2. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in**

the response of yeast cells to environmental changes. *Mol Biol Cell* 2000, **11:**4241-4257.
3. Yun CW, Ferea T, Rashford J, Ardon O, Brown PO, Botstein D, Kaplan J, Philpott CC: **Desferrioxamine-mediated iron uptake in *Saccharomyces cerevisiae*. Evidence for two pathways of iron uptake.** *J Biol Chem* 2000, **275:**10709-10715.
4. Lyons TJ, Gasch AP, Gaither LA, Botstein D, Brown PO, Eide DJ: **Genome-wide characterization of the Zap1p zinc-responsive regulon in yeast.** *Proc Natl Acad Sci USA* 2000, **97:**7957-7962.
5. Ogawa N, DeRisi J, Brown PO: **New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis.** *Mol Biol Cell* 2000, **11:**4309-4321.
6. Travers KJ, Patil CK, Wodicka L, Lockhart DJ, Weissman JS, Walter P: **Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation.** *Cell* 2000, **101:**249-258.
7. MacQueen J: **Some methods for classification and analysis of multivariate observations.** In *Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics and Probability*. Edited by Le Cam L, Neyman J. Berkeley: University of California, Berkeley Press; 1967: 281-297.
8. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22:**281-285.
9. *Fuzzy models for Pattern Recognition*. Edited by Bezdek JC, Pal SK. New York: IEEE Press; 1992.
10. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data.** *J Comput Biol* 2000, **7:**601-620.
11. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N: **Revealing modular organization in the yeast transcriptional network.** *Nat Genet* 2002, **31:**370-377.
12. Bezdek JC: *Fuzzy Mathematics in Pattern Classification.* Ithaca, NY: Cornell University; 1973.
13. Gath I, Geva AB: **Unsupervised optimal fuzzy clustering.** *Trans Pattern Analysis Machine Intell* 1989, **11:**773-781.
14. **FuzzyK** [http://rana.lbl.gov/FuzzyK]
15. Aldenderfer MS, Blashfield RK: **Cluster analysis.** In *Quantitative Applications in the Social Sciences*, *Vol. 88*. Edited by Lewis-Beck MS. Newbury Park: Sage; 1984.
16. Fraley C, Raftery AE: **How many clusters? Which clustering method? Answers via model-based cluster analysis.** *Techn Rep* 1998, **329:**1-19.
17. Dudoit S, Fridlyand J: **A prediction-based resampling method for estimating the number of clusters in a dataset.** *Genome Biol* 2002, **3:**1-21.
18. Raychaudhuri S, Stuart JM, Altman RB: **Principal components analysis to summarize microarray experiments: application to sporulation time series.** *Pac Symp Biocomput* 2000, 455-466.
19. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling.** *Proc Natl Acad Sci USA* 2000, **97:**10101-10106.
20. Gasch AP, Huang M, Metzner S, Botstein D, Elledge SJ, Brown PO: **Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p.** *Mol Biol Cell* 2001, **12:**2987-3003.
21. ***Saccharomyces* Genome Database** [http://genome-www.stanford.edu/Saccharomyces/]
22. Nielsen H, Engelbrecht J, Brunak S, von Heijne G: **A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.** *Int J Neural Sys* 1997, **8:**581-599.
23. Gething MJ: **Role and regulation of the ER chaperone BiP.** *Semin Cell Dev Biol* 1999, **10:**465-472.
24. Jamsa E, Simonen M, Makarow M: **Selective retention of secretory proteins in the yeast endoplasmic reticulum by treatment of cells with a reducing agent.** *Yeast* 1994, **10:**355-370.
25. Cox JS, Walter P: **A novel mechanism for regulating activity of a transcription factor that controls the unfolded protein response.** *Cell* 1996, **87:**391-404.
26. Mori K, Kawahara T, Yoshida H, Yanagi H, Yura T: **Signalling from endoplasmic reticulum to nucleus: transcription factor with a basic-leucine zipper motif is required for the unfolded protein- response pathway.** *Genes Cells* 1996, **1:**803-817.
27. Nikawa J, Akiyoshi M, Hirata S, Fukuda T: ***Saccharomyces cerevisiae IRE2/HAC1* is involved in IRE1-mediated *KAR2* expression.** *Nucleic Acids Res* 1996, **24:**4222-4226.
28. Chapman R, Sidrauski C, Walter P: **Intracellular signaling from the endoplasmic reticulum to the nucleus.** *Annu Rev Cell Dev Biol* 1998, **14:**459-485.
29. Cox JS, Shamu CE, Walter P: **Transcriptional induction of genes encoding endoplasmic reticulum resident proteins requires a transmembrane protein kinase.** *Cell* 1993, **73:**1197-1206.
30. Mori K, Ma W, Gething MJ, Sambrook J: **A transmembrane protein with a cdc2$^+$/CDC28-related kinase activity is required for signaling from the ER to the nucleus.** *Cell* 1993, **74:**743-756.
31. Oka M, Kimata Y, Mori K, Kohno K: ***Saccharomyces cerevisiae KAR2* (BiP) gene expression is induced by loss of cytosolic HSP70/Ssa1p through a heat shock element-mediated pathway.** *J Biochem (Tokyo)* 1997, **121:**578-584.
32. Arndt KT, Styles C, Fink GR: **Multiple global regulators control HIS4 transcription in yeast.** *Science* 1987, **237:**874-880.
33. Daignan-Fornier B, Fink GR: **Coregulation of purine and histidine biosynthesis by the transcriptional activators *BAS1* and *BAS2*.** *Proc Natl Acad Sci USA* 1992, **89:**6746-6750.
34. Jia MH, Larossa RA, Lee JM, Rafalski A, Derose E, Gonye G, Xue Z: **Global expression profiling of yeast treated with an inhibitor of amino acid biosynthesis, sulfometuron methyl.** *Physiol Genomics* 2000, **3:**83-92.
35. Natarajan K, Meyer MR, Jackson BM, Slade D, Roberts C, Hinnebusch AG, Marton MJ: **Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast.** *Mol Cell Biol* 2001, **21:**4347-4368.
36. Cai M, Davis RW: **Yeast centromere binding protein CBF1, of the helix-loop-helix protein family, is required for chromosome stability and methionine prototrophy.** *Cell* 1990, **61:**437-446.
37. Thomas D, Jacquemin I, Surdin-Kerjan Y: **MET4, a leucine zipper protein, and centromere-binding factor 1 are both required for transcriptional activation of sulfur metabolism in *Saccharomyces cerevisiae*.** *Mol Cell Biol* 1992, **12:**1719-1727.
38. Blaiseau PL, Isnard AD, Surdin-Kerjan Y, Thomas D: **Met31p and Met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism.** *Mol Cell Biol* 1997, **17:**3640-3648.
39. Blaiseau PL, Thomas D: **Multiple transcriptional activation complexes tether the yeast activator Met4 to DNA.** *EMBO J* 1998, **17:**6327-6336.
40. Nehlin JO, Ronne H: **Yeast MIG1 repressor is related to the mammalian early growth response and Wilms' tumour finger proteins.** *EMBO J* 1990, **9:**2891-2898.
41. de Winde JH, Grivell LA: **Global regulation of mitochondrial biogenesis in *Saccharomyces cerevisiae*.** *Prog Nucleic Acid Res Mol Biol* 1993, **46:**51-91.
42. Gancedo JM: **Yeast carbon catabolite repression.** *Microbiol Mol Biol Rev* 1998, **62:**334-361.
43. Gorner W, Durchschlag E, Wolf J, Brown EL, Ammerer G, Ruis H, Schuller C: **Acute glucose starvation activates the nuclear localization signal of a stress-specific yeast transcription factor.** *EMBO J* 2002, **21:**135-144.
44. Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, Lee TI, True HL, Lander ES, Young RA: **Remodeling of yeast genome expression in response to environmental changes.** *Mol Biol Cell* 2001, **12:**323-337.
45. Mager WH, Planta RJ: **Multifunctional DNA-binding proteins mediate concerted transcription activation of yeast ribosomal protein genes.** *Biochim Biophys Acta* 1990, **1050:**351-355.
46. Moehle CM, Hinnebusch AG: **Association of RAP1 binding sites with stringent control of ribosomal protein gene transcription in *Saccharomyces cerevisiae*.** *Mol Cell Biol* 1991, **11:**2723-2735.
47. Hughes JD, Estep PW, Tavazoie S, Church GM: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*.** *J Mol Biol* 2000, **296:**1205-1214.
48. Bussemaker HJ, Li H, Siggia ED: **Regulatory element detection using correlation with expression.** *Nat Genet* 2001, **27:**167-171.
49. Fazzio TG, Kooperberg C, Goldmark JP, Neal C, Basom R, Delrow J, Tsukiyama T: **Widespread collaboration of Isw2 and Sin3-Rpd3 chromatin remodeling complexes in transcriptional repression.** *Mol Cell Biol* 2001, **21:**6450-6460.
50. Kurdistani SK, Robyr D, Tavazoie S, Grunstein M: **Genome-wide binding map of the histone deacetylase Rpd3 in yeast.** *Nat Genet* 2002, **31:**248-254.

51. Andrews BJ, Herskowitz I: **The yeast SWI4 protein contains a motif present in developmental regulators and is part of a complex involved in cell-cycle-dependent transcription.** *Nature* 1989, **342:**830-833.
52. Andrews BJ, Herskowitz I: **Identification of a DNA binding factor involved in cell-cycle control of the yeast HO gene.** *Cell* 1989, **57:**21-29.
53. Koch C, Moll T, Neuberg M, Ahorn H, Nasmyth K: **A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase.** *Science* 1993, **261:**1551-1557.
54. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: **Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.** *Nature* 2001, **409:**533-538.
55. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.** *Mol Biol Cell* 1998, **9:**3273-3297.
56. Bailey TB, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. Menlo Park, CA: AAAI Press; 1994.
57. Kosower NS, Kosower EM: **Diamide: an oxidant probe for thiols.** *Methods Enzymol* 1995, **251:**123-133.
58. Shenton D, Perrone G, Quinn KA, Dawes IW, Grant CM: **Regulation of protein S-thiolation by glutaredoxin 5 in the yeast *Saccharomyces cerevisiae*.** *J Biol Chem* 2002, **277:**16853-16859.
59. Tortorella D, Story CM, Huppa JB, Wiertz EJ, Jones TR, Bacik I, Bennink JR, Yewdell JW, Ploegh HL: **Dislocation of type I membrane proteins from the ER to the cytosol is sensitive to changes in redox potential.** *J Cell Biol* 1998, **142:**365-376.
60. Salgueiro MJ, Krebs N, Zubillaga MB, Weill R, Postaire E, Lysionek AE, Caro RA, De Paoli T, Hager A, Boccio J: **Zinc and diabetes mellitus: is there a need of zinc supplementation in diabetes mellitus patients?** *Biol Trace Element Res* 2001, **81:**215-228.
61. Hebert DN, Simons JF, Peterson JR, Helenius A: **Calnexin, calreticulin, and Bip/Kar2p in protein folding.** *Cold Spring Harb Symp Quant Biol* 1995, **60:**405-415.
62. Baksh S, Spamer C, Heilmann C, Michalak M: **Identification of the Zn^{2+} binding region in calreticulin.** *FEBS Lett* 1995, **376:**53-57.
63. Cox LS: **Multiple pathways control cell growth and transformation: overlapping and independent activities of p53 and p21Cip1/WAF1/Sdi1.** *J Pathol* 1997, **183:**134-140.
64. Pritts T, Hungness E, Wang Q, Robb B, Hershko D, Hasselgren PO: **Mucosal and enterocyte IL-6 production during sepsis and endotoxemia - role of transcription factors and regulation by the stress response.** *Am J Surg* 2002, **183:**372-383.
65. Thomas D, Surdin-Kerjan Y: **Metabolism of sulfur amino acids in *Saccharomyces cerevisiae*.** *Microbiol Mol Biol Rev* 1997, **61:**503-532.
66. **Eisen lab** [http://rana.lbl.gov]
67. **Cluster and TreeView Manual** [http://rana.lbl.gov/manuals/ClusterTreeView.pdf]
68. **Regulatory Sequence Analysis Tools** [http://rsat.ulb.ac.be/rsat/]
69. van Helden J, Andre B, Collado-Vides J: **A web site for the computational analysis of yeast regulatory sequences.** *Yeast* 2000, **16:**177-187.