Subsystem Identification Through Dimensionality Reduction of Large-Scale Gene Expression Data

Philip M. Kim[1] and Bruce Tidor[2,3,4]

[1]*Department of Chemistry,* [2]*Biological Engineering Division,* [3]*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

The availability of parallel, high-throughput biological experiments that simultaneously monitor thousands of cellular observables provides an opportunity for investigating cellular behavior in a highly quantitative manner at multiple levels of resolution. One challenge to more fully exploit new experimental advances is the need to develop algorithms to provide an analysis at each of the relevant levels of detail. Here, the data analysis method non-negative matrix factorization (NMF) has been applied to the analysis of gene array experiments. Whereas current algorithms identify relationships on the basis of large-scale similarity between expression patterns, NMF is a recently developed machine learning technique capable of recognizing similarity between subportions of the data corresponding to localized features in expression space. A large data set consisting of 300 genome-wide expression measurements of yeast was used as sample data to illustrate the performance of the new approach. Local features detected are shown to map well to functional cellular subsystems. Functional relationships predicted by the new analysis are compared with those predicted using standard approaches; validation using bioinformatic databases suggests predictions using the new approach may be up to twice as accurate as some conventional approaches.

[Supplemental material is available online at www.genome.org.]

Gene-expression microarrays are a recently developed technology that allows genome-wide measurement of RNA expression levels in a highly quantitative fashion (Fodor et al. 1993; Schena et al. 1995; Granjeaud et al. 1999). Studies with microarrays generally produce large two-dimensional data sets (e.g., simultaneous monitoring of thousands of genes measured in up to hundreds of different experiments; Cho et al. 1998; Chu et al. 1998; Spellman et al. 1998; Iyer et al. 1999; Hughes et al. 2000; Kim et al. 2001). The promise of this type of highly parallel and quantitative data is that they contain detailed and subtle information about relationships among cellular, biochemical, and genetic components that underlie the behavior of cells; the difficulty is that current approaches lead to data that are somewhat noisy (Coller et al. 2000; Brown et al. 2001; Li and Wong 2001), and the development of methods for exploring and extracting relationships within the data is still in its infancy.

The collection, processing, and analysis of microarray data present many challenges. Appropriate treatment of noise and systematic error is necessary to ensure that further analysis is not clouded by data inaccuracy, and some approaches have been proposed (e.g., Brown et al. 2001; Li and Wong 2001; Broet et al. 2002). Methods of analysis must be developed that answer particular and relevant questions. Often, these questions involve seeking and identifying patterns of similarity (correlation or anticorrelation) within the data. An array of methods capable of recognizing different types of similarity and similarity at different levels of resolution is needed. Moreover, the development of approaches to test individual hypotheses given a particular set of data and to more fully incorporate pre-existing models (Getz et al. 2000; Hartemink et al. 2001; Ideker et al. 2001; Tanay and Shamir 2001) or other sources of information in the analysis is an important research area (Golub et al. 1999; Bittner et al. 2000; Brown et al. 2000).

One productive use of expression data is to propose and to study relationships between genetic, cellular, or environmental components. Examples include the elucidation of metabolic (DeRisi et al. 1997; Ferea et al. 1999) or regulatory (Holstege et al. 1998; Tavazoie et al. 1999; Ren et al. 2000) networks. The standard methodology involves clustering of expression patterns on the basis of similarity (Chu et al. 1998; Eisen et al. 1998; Alon et al. 1999; Heyer et al. 1999; Tamayo et al. 1999; Tavazoie et al. 1999; Sherlock 2000; Zhu et al. 2002). The main assumption generally applied is that similar gene expression profiles imply related function. There are other techniques, many of which come from the machine-learning community, capable of detecting similarity or partially repeated patterns in large data sets. In principle, these techniques provide alternative approaches for recognizing potential relationships within large biological data samples, including expression arrays, that may complement existing methods. Here, one such machine-learning algorithm, non-negative matrix factorization (NMF), has been applied to the analysis of microarray data. One characteristic of NMF is that, using dimensionality reduction, it is capable of identifying patterns that exist in only a subset of the data (Lee and Seung 1999). For example, the application of clustering to recognize experimental conditions with similar patterns of gene expression focuses attention on conditions for which similarity extends across all genes. Although more recent techniques account for the combinatorial nature of gene regulation (Gasch and Eisen 2002; Zhu et al. 2002), they still focus on global patterns of similarity. Another data analysis approach, singular value decomposition (SVD), also bases its description of

[4]Corresponding author.
E-MAIL tidor@mit.edu; FAX (617) 252-1816.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.903503.

13:1706–1718 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

the underlying data on global relationships that extend across essentially all of the data and has been applied recently to microarray data (Misra et al. 2002). In contrast, NMF recognizes sets of experimental conditions in which smaller sets of genes behave in a strongly correlated fashion. Thus, whereas other analysis methods examine global patterns in search of similarity and correlation, NMF is capable of finding smaller, more localized patterns as well as global patterns. Such an approach might be particularly useful in identifying biological subsystems (i.e., sets of genes that function in concert in a relatively tightly regulated manner) and might be an especially sensitive means for detecting functional genetic relationships.

Here, the potential usefulness of NMF for the analysis of high-dimensional biological data was evaluated using a publicly available compendium microarray data set for *Saccharomyces cerevisiae*, in which 6316 ORFs were monitored in each of 300 experiments (Hughes et al. 2000). Most of the experiments (276 of the 300) corresponded to deletion mutants of individual genes. In addition, 13 involved mutants with individual genes overexpressed using tetracycline-regulated alleles, and 11 involved wild-type cells treated with specific drugs. This data set spans a relatively wide set of significant cellular perturbations. The size of the data set is large by current standards, which presents a challenge for computational approaches but also an opportunity to find patterns in what appears to be a particularly rich set of experiments. Analysis using NMF suggested that reduction of the data to a 50-dimensional subspace is appropriate. The lower dimensional subspace was capable of reconstructing the original data to high fidelity. The 50 vectors describing the subspace were relatively insensitive to moderate amounts of noise added to the original data set. The vectors described the local feature space detected by NMF and showed that each set of features was dominated by a few functional categories, indicating that they represent a grouping of genetic components on the basis of cellular function. Individual pairwise functional relationships were scored on the basis of standard approaches and, alternatively, using the similarity as measured by NMF. Scoring the relationships using the Munich Information Center for Protein Sequences functional categories (MIPS categories; http://mips.gsf.de/; Mewes et al. 2000) and the Yeast Proteome Database (YPD; Proteome, Inc.; http://www.incyte.com/; Costanzo et al. 2001) indicated that the new approach is significantly more reliable at predicting relationships than standard approaches. NMF appears to be a promising methodology, complementary to current approaches, for the analysis of high-dimensional biological data.

RESULTS

The compendium data set contained expression patterns monitored for 6316 *S. cerevisiae* genes in 300 experiments involving a variety of strains and conditions. The expression of each gene in each experiment was represented as a ratio of the expression in the experiment to that in a control experiment of wild type grown under standard conditions. Genes whose expression in the control was not measurable, were removed from the data set to prevent division by zero, leaving 5346 genes, and the natural logarithm of each ratio was taken. Data analysis involved using NMF to reduce the dimensionality of the data and to extract common features repeated in correlated fashion throughout the data (see Methods). These common feature elements were represented as basis vectors resulting from the technique. In typical usage, each basis vector represented an experiment, in that it contained a relative expression for each gene comprising the feature represented.

Selection of NMF Dimensionality

An essential feature of the NMF approach is that it reduces the data set from its full dimensionality (original data space) to a lower dimensional NMF space. Initial calculations were performed to select an appropriate size for the lower dimensional NMF space. Trial calculations carried out with NMF dimension of size 10–80 suggested that 50 represented a good compromise that provided an adequate reconstruction of the experimental data while giving basis vectors that appeared to recognize repetitive features. The RMS error between the origi-

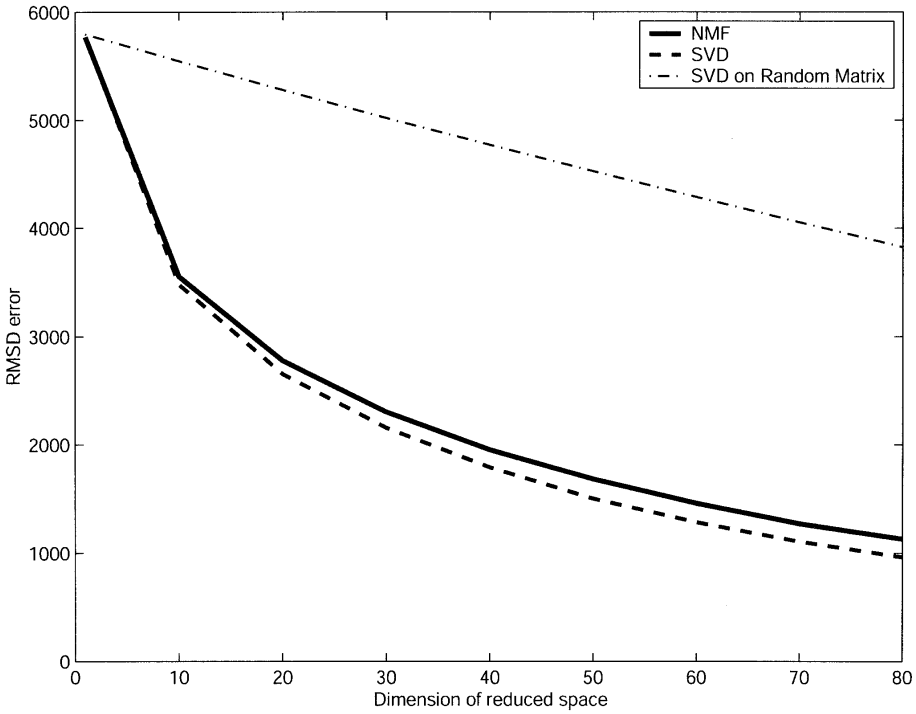

Figure 1 The RMS error of NMF and SVD factorizations of the original data as a function of the number of dimensions in the reduced space. For comparison, SVD factorization was also carried out on a random matrix based on the data matrix. The results show that NMF is nearly as good as SVD at reproducing the original data for any dimensionality, and that near a dimensionality of about 50 the marginal increase (slope) in NMF's ability to describe the original data is similar to SVD's ability to match random (unstructured) data. Thus, an NMF dimensionality of 50 is appropriate to describe the structure in the data.

nal and NMF reconstructed data is shown as a function of the size (dimensionality) of the NMF space in Figure 1. Also shown in Figure 1 is the RMS error for singular value decomposition (SVD), which is another matrix factorization approach that is guaranteed to produce the minimum error for a given dimensionality (but does not generally extract localized features from complex data sets). SVD was applied both to the actual data matrix and to a random matrix of values selected from a gaussian distribution with the same mean and variance as the data matrix and subject to the non-negativity constraint. The close similarity between the error for NMF and SVD on the actual data indicated that the computational procedures used for NMF were effective (details given in Methods). The random matrix could be viewed as one without correlated features to be detected through factorization; the slope of the RMS error plot for this matrix represents the added ability to reproduce unstructured data with additional basis vectors. Below a dimensionality of 50, the NMF factorization curve had a steeper slope than the random matrix line, which indicated improvements due to capturing organization and structure within the data. This further justified the choice of 50 for the NMF dimension. Interestingly, a previous study using expression arrays to study yeast also found an inherent dimensionality of 50 (Alter et al. 2000).

Basis vectors (basis experiments) obtained from NMF factorization with a dimensionality of 50 were sparse and reproducible. One measure of sparsity is the fraction of non-zero entries per basis vector, which averaged 5% over the 50 vectors. The factorization produced somewhat different results each time it was started from a different random starting point. When the basis vectors from different factorizations using the same dimensionality were compared, the correlation coefficient was found to be >0.9 between pairs. This indicates that results of NMF are robust with respect to the mathematical procedures used here to perform the calculations. The RMS error of the reconstructed data (through NMF di-

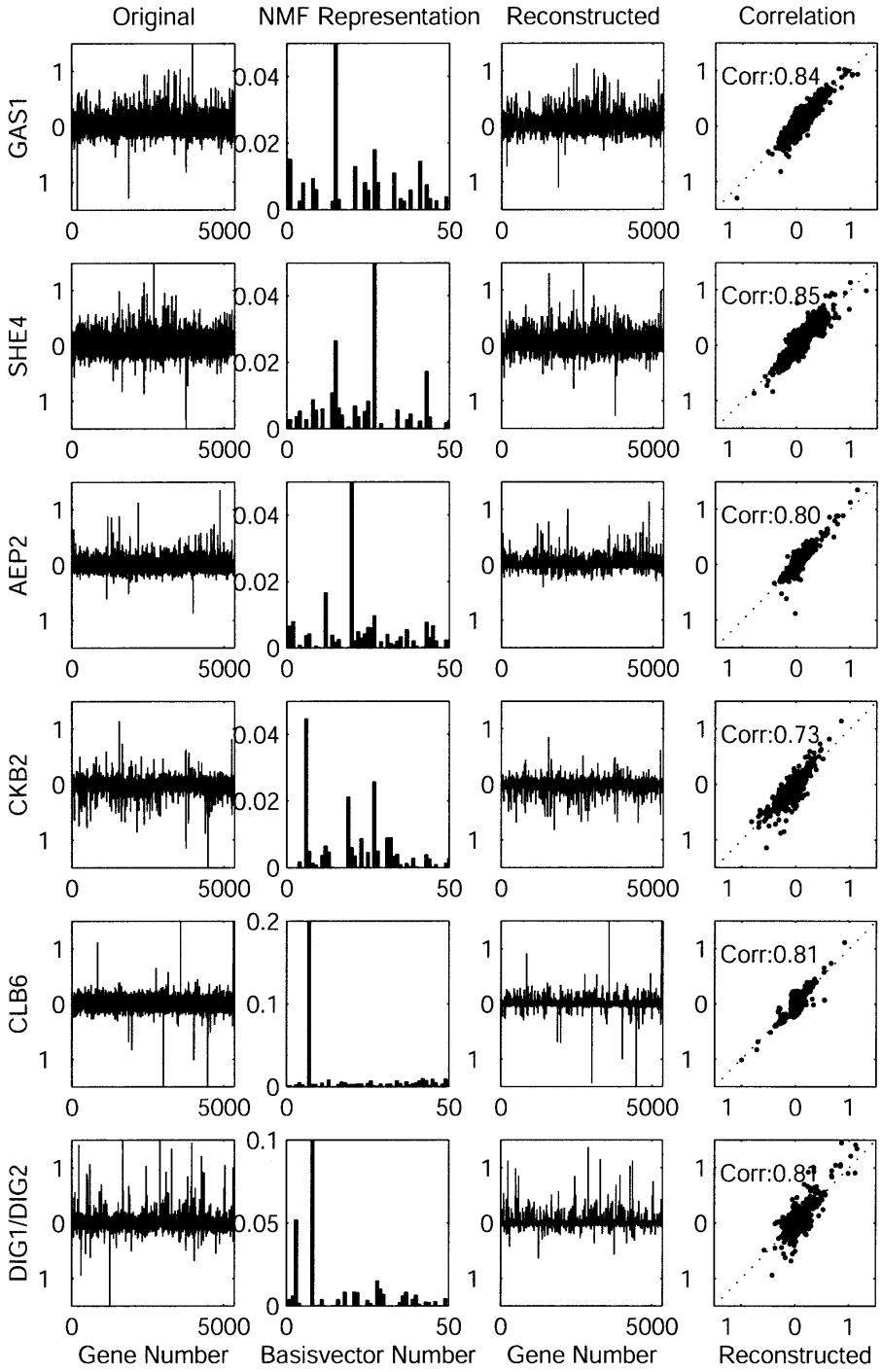

Figure 2 Representation of gene expression data in full and NMF-reduced spaces. (*Left* column) The original data (log-ratio) is shown for 6 individual experiments in the space of 5346 genes, in the second column from *left*, the 50-dimensional NMF representation is shown. In the third column from *left*, the reconstruction from the NMF representation back to the original space (using $W \cdot H$) is shown. (*Right* column) The log-ratios of the original (y-axis) are plotted against the log-ratios of the reconstruction from the NMF representation back to the original experimental space (x-axis). The data show that the NMF reduction is capable of regenerating the experiments to relatively high fidelity, and that the NMF representation of an experiment is often dominated by one or a small number of features (basis vectors).

mensional reduction) compared with the original data was only about 7.8% of the RMS error of a random permutation of the original data, in which the experiments (columns) of the original data matrix were permuted. Figure 2 illustrates six examples of expression experiments in the original gene expression space, in the 50-dimensional NMF space, and reconstructed from the NMF space back into the original space. This shows the ability of the dimensionality reduction to still capture many of the details of the original data. Also, it is demonstrated that most experiments are dominated by the combination of only a few important basis vectors. They correspond to similarities across many, but not all genes.

To examine the robustness of the algorithm to noise, gaussian noise was added to the original data to produce corrupted data vectors. Table 1 lists the average correlation between results of the analysis performed on the original and corrupted data. Gaussian noise was added in progressively larger increments of the standard deviation of the data. As a recent model that captures the physical processes underlying microarray measurements indicates, the ratio of the cDNA distributions can be approximated with a log-normal distribution (K.H. Duggar, T. Ideker, D.A. Lauffenburger, and P.K. Sorger, in prep.). These results suggest that gaussian noise is appropriate to apply to the log of the ratios. At low noise (0.2 times the standard deviation), there was very little change in the results. The correlation of NMF vectors was better than 0.90, as was that for the data reconstructed from the dimensionality reduction. This is not surprising, because the original data vectors and corrupted vectors also showed a correlation coefficient of >0.90. However, when adding more noise (equal to the full standard deviation), both the NMF basis vectors, as well as the reconstructed data, were still very similar after adding noise (correlation of better than 0.80), whereas the original data was changed substantially more (correlation of 0.57). This fact shows the high robustness of NMF to noise in the data, and suggests that NMF might be useful as a noise-reduction filter in certain applications.

Annotation of Basis Vectors

Each of the 50 basis vectors (basis experiments) contained many genes with zero expression and others with non-zero expression. Because of the sparsification procedure applied (see Methods), 95% of the entries across all basis vectors were constrained to be exactly zero. The genes with non-zero expression were used to assign sets of functional categories to basis vectors using the MIPS classification scheme (see Methods; Mewes et al. 2000), and the results are listed in Table 2. We emphasize that this analysis of NMF basis vectors in terms of functional categories was done to understand the nature of the basis vectors and not to make predictions about relatedness of genes. When predictions of functional relatedness are made (next subsection), they are made on the basis of changes of expression levels in experiments using strains deleted for each of the two genes in question. Changes measured in the NMF space are compared with changes measured in other spaces.

Table 1. Robustness of NMF Basis Vectors to Noise

Noise added	NMF basis vectors	Reconstructed data	Original data
0.2	0.933	0.930	0.943
0.5	0.879	0.893	0.781
1	0.865	0.816	0.573
5	0.368	0.313	0.159

Gaussian noise was added to the original data and was quantified as a multiplier of the standard deviation of the original data set. (NMF basis vectors) The average correlation of basis vectors from the original data to the basis vectors from original data with added noise. (Reconstructed data) The average correlation of the reconstructed data from the basis vectors with and without noise. (Original data) Average correlation of the original data to the data with added noise.

Each basis vector appeared to be dominated by only a few functional categories, with some categories showing increased and others decreased expression relative to wild-type, untreated cells. Basis vector 17, for example, showed increased expression of genes associated with amino-acid metabolism and metabolism of energy reserves together with decreased expression of genes involved in rRNA transcription. Basis vector 20 showed increased expression of genes involved in ion transport, homeostasis of cations, ribosomal function, and mitochondrial organization with decreased expression of genes for amino-acid metabolism, (other) ribosomal proteins, translation, and organization of cytoplasm. Basis vector 9 showed increased expression of genes associated with carbon compound (C-compound) and carbohydrate metabolism and transporters as well as metabolism of energy reserves, and at the same time decreased expression of amino-acid metabolism genes. In some cases, specific metabolic pathways could be seen in the basis vectors. For instance, fatty acid oxidation was up-regulated in basis vector 42. Most elements of the TCA-cycle were up-regulated in basis vector 43. Furthermore, this basis vector, which seemed mostly responsible for energy metabolism, contained all but two of the genes involved in the pentose-phosphate shunt. Of these two genes, one is a transketolase that is highly homologous to another transketolase found in basis vector 43, and the other is the ribose-5-phosphate ketol isomerase. In 14 of the basis vectors, no single MIPS category was significantly enriched, which is partly due to the lack of sparsity (i.e., too many genes occur in a basis vector, therefore, no single category was significant), and partly due to an abundance of as yet uncategorized genes.

Independent of the classification scheme proposed by MIPS, the occurrence of well-characterized gene groups was examined in basis vectors. The processed data set contained nine histone genes, which were all present together in basis vector 1. This enrichment was >5σ higher than what would occur by chance. Aside from histone genes, basis vector 1 was also strongly enriched in ribosomal genes, genes related to translation, and genes involved in amino-acid and nitrogen metabolism. Similarly, the data set contained 109 ribosomal genes, of which 70 appeared in basis vector 1 and 52 in basis vector 43. Basis vector 43 was involved in energy metabolism, stress response, and rRNA transcription. The enrichment of 70 ribosomal genes in basis vector 1 was 26σ higher than would occur by chance. Between basis vectors 1 and 43, all but 17 ribosomal genes were found.

Next, the occurrence of genes in both the *GAL4* and the *STE12* pathway was examined. These pathways were recently studied extensively by Ren et al. (2000). No deletion mutant of any of the genes involved in the *GAL4* pathways was present in the compendium data set; therefore, no significant enrichment of those genes might be expected. However, of the nine genes present in the data, five were enriched in basis vector 9. This enrichment was 5σ higher than would be ex-

Table 2. Annotation of 12 of the 50 NMF Basis Vectors Based on the MIPS Functional Categories

1	+1 amino-acid metabolism (204 ORFs) [82] +2 nitrogen and sulphur metabolism (74 ORFs) [27] +81 stress response (169 ORFs) [43] −34 ribosomal proteins (206 ORFs) [98] −35 translation (62 ORFs) [22] −92 organization of cytoplasm (557 ORFs) [163]	687
3	+1 amino-acid metabolism (204 ORFs) [21]	141
4	+81 stress response (169 ORFs) [9]	53
8	+21 pheromone response, mating-type det., sex-spec. proteins (159 ORFs) [24] −4 phosphate metabolism (31 ORFs) [3]	211
9	+5 C-compound and carbohydrate metabolism (413 ORFs) [115] +15 metabolism of energy reserves (glycogen, trehalose) (37 ORFs) [21] +47 C-compound and carbohydrate transporters (46 ORFs) [23] −1 amino-acid metabolism (204 ORFs) [53]	1007
17	+1 amino-acid metabolism (204 ORFs) [40] +15 metabolism of energy reserves (glycogen, trehalose) (37 ORFs) [14] −29 rRNA transcription (104 ORFs) [17]	203
19	+13 respiration (85 ORFs) [18] +100 mitochondrial organization (364 ORFs) [27] −1 amino-acid metabolism (204 ORFs) [18]	155
20	+34 ribosomal proteins (206 ORFs) [29] +46 ion transporters (76 ORFs) [13] +88 homeostasis of cations (112 ORFs) [17] +100 mitochondrial organization (364 ORFs) [53] −1 amino-acid metabolism (204 ORFs) [18] −34 ribosomal proteins (206 ORFs) [24] −35 translation (62 ORFs) [7] −92 organization of cytoplasm (557 ORFs) [40]	207
23	+11 tricarboxylic-acid pathway (23 ORFs) [4] +15 metabolism of energy reserves (glycogen, trehalose) (37 ORFs) [6] −4 phosphate metabolism (31 ORFs) [3]	128
36	+29 rRNA transcription (104 ORFs) [41] −5 C-compound and carbohydrate metabolism (413 ORFs) [77] −11 tricarboxylic-acid pathway (23 ORFs) [10] −15 metabolism of energy reserves (glycogen, trehalose) (37 ORFs) [14]	557
42	+1 amino-acid metabolism (204 ORFs) [40] +6 lipid, fatty-acid and isoprenoid metabolism (210 ORFs) [21] +81 stress response (169 ORFs) [22] −21 pheromone response, mating-type det., sex-spec. proteins (159 ORFs) [11]	199
43	+5 C-compound and carbohydrate metabolism (413 ORFs) [126] +10 pentose-phosphate pathway (9 ORFs) [7] +11 tricarboxylic-acid pathway (23 ORFs) [17] +81 stress response (169 ORFs) [64] −29 rRNA transcription (104 ORFs) [58] −34 ribosomal proteins (206 ORFs) [87]	1111

Each annotation includes a plus or minus sign (indicating whether expression is enhanced or decreased compared with control experiments), an integer number indexing the MIPS category, the name of the MIPS category, the number of ORFs belonging to the MIPS category, and the number of genes in the basis vector belonging to the MIPS category (in square brackets). The third column indicates the number of genes that are non-zero in this basis vector. The full set of 50 basis vectors is provided as supplementary information.

pected by chance. Basis vector 9 was also involved in C-compound and carbohydrate metabolism. It seemed that basis vector 9 was responsible for a broad range of functions relating to carbohydrate metabolism, the utilization of galactose being a subset of those.

There was a deletion mutant of *STE12* in the data, along with several mutants of related genes, including *FUS3*, *KSS1*, and *STE5*. A total of 25 genes, forming a subset of those identified by Ren et al. (2000) as members of the *STE12* pathway, were present in the processed data. A majority, 16 of the 25, were present in basis vector 8 (a significance of 16σ higher than expected by chance). The genes included *PRM1* (linked to membrane biosynthesis), *FIG2*, *AGA1*, *FUS1* (cell fusion), *GIC2* (mating projection formation), *CIK1*, *KAR2* (nuclear fusion), *FUS3*, *STE12*, and *HYM1* (mating signaling) along with other genes of yet unknown relevance (*YOR0343C*, *PEP1*, *SCH9*, *YIL036C*, *YIL083C*, *YOL155C*). It should be noted that all genes to which Ste12p binds (as identified by Ren et al. 2000) before and after α factor addition were included in this list (a total of 8 genes; 17 additional genes that were in the preprocessed data set were shown to bind Ste12p only after α factor addition). Basis vector 8 also included large contributions from genes represented in the MIPS database as involved in mating signaling and pheromone response, indicating related cellular functions for the genes dominating this basis vector. Another MIPS category that was found to be enriched in basis vector 8 with 4.5σ (just below the cutoff implemented of 5σ) was membrane biosynthesis, which is consistent with the appearance of *PRM1*, classified as an effector in membrane biosynthesis.

Basis vector 8 (the mating basis vector) was then examined more closely, and the function of all of its member genes examined using information from the Yeast Proteome Database (YPD), constructed by Proteome, Inc. (http://www.incyte.com/; Costanzo et al. 2001). The YPD is a compilation of published results of yeast genes (*S. cerevisiae*) and their functions, including functional relationships reported in the literature. Aside from the 16 genes described above, it contained 15 other genes involved in mating or pheromone response (*TEC1*, *KAR4*, *PRM3*, *PGU1*, *YLR042C*, *DDR48*, *PRM5*, *SAG1*, *HAP4*, *SST2*, *MSG5*, *AGA1*, *PRM4*, *SAG1*, *KSS1*), 6 of which (underlined) were annotated in YPD as directly induced by *STE12*. Furthermore, this vector contained 10 genes (*ECM18*, *SPI1*, *CHS7*, *GFA1*, *KTR2*, *SCW10*, *WSC3*, *STR2*, *GSC2*, *PHD1*) involved in cell-wall or cell-membrane biosynthesis or maintenance. Among its other members were several genes involved in carbohydrate metabolism (*GLK1*, *SOL4*, *GPH1*, *GLC3*), heat shock or stress response (*HSP26*, *HSP30*, *PRY2*, *DDR48*), and many ORFs of yet unknown function (*YDR124W*, *YDR537C*, *PTI1*, *YGR250C*, *YHR213W*, *YIL060W*, *YIL082W*, *YIL083C*, *YJR026W*, *YJR027W*, *YJR028W*, *SRL3*, *YLR177W*, *YLR334C*, *YLR422W*, *YOL106W*, *YOR296W*, *SVS1*), as well as a few genes of other functionality (*ADR1*, *BNA1*, *FRE2*). Besides examining the genes that contribute strongly to the basis vector, it is informative to examine which of the 300 experiments in the compendium were described by use of a large contribution from this basis vector. Basis vector 8 was used mostly to de-

scribe experiments of deletion mutants of *DIG1/DIG2* (double deletion; *DIG1* is a known *STE12* repressor), *DIG1* (single deletion), and *FUS3* (linked to mating and pheromone response). Note that the data set did not contain a *DIG2* single-deletion mutant.

Prediction of Functional Relationships

The compendium data set analyzed here was dominated by measurements of gene expression in deletion strains of yeast compared with wild type (276 of 300 experiments). Of the remaining experiments, 13 were measurements of single-gene overexpression relative to wild type. Thus, all but 11 experiments involved direct manipulation of a single gene in a common background. (The 11 exceptions involved measurements of wild-type yeast treated with a single drug relative to untreated wild type.) Experiments showing similar (correlated or anti-correlated) changes in gene expression at some level might be expected to be functionally related. In particular, the two genes deleted in the two experiments could be expected to be part of the same or related cellular function. To test this hypothesis, predictions of functional relationships were made and scored against available database information. Moreover, predictions based on correlations in the entire gene space were compared with those from dimensionally reduced spaces, such as that produced by NMF, to understand whether dimensionality reduction can enhance the detection of known genetic relationships. One difficulty with any approach of this type is that available database information is likely to be incomplete and may be partially inaccurate. Nevertheless, a method's ability to recapitulate current knowledge is a good indicator of its ability to predict new relationships. Thus, the score these methods achieve in validated functional relationships should only be interpreted relative to each other, as many true functional relationships may be missing from current databases.

Predictions of functional relationships were made using the pairwise correlations between experiments measured in each of six spaces—the original data space, the 50-dimensional NMF space, and four other 50-dimensional spaces chosen for comparison. The six spaces are (1) the original space in which the data was collected, corresponding to 5346 genes used in the analysis, (2) the 50-dimensional space resulting from NMF data reduction, (3) the 50-dimensional space spanned by 50 genes whose expression varied the most across the 300 experiments, (4) the 50-dimensional space explaining the largest variation in the experimental data as found by SVD, (5) the 50-dimensional space resulting from NMF data reduction without applying the sparsification procedure, and (6) the space spanned by the eigenvectors from SVD that have been subjected to the same sparsification procedure as NMF in (2). In addition, comparison was made to the average validity of predictions made from k-means clustering with 50 clusters. Note that the comparison with clustering is not completely fair, as here we are testing for pairwise relationships between genes, whereas clustering finds groupwise relationships. To compare k-means clustering, we treat each pair of experiments within a given cluster as related. For each case, the pairwise correlations were sorted by magnitude, with the higher magnitude correlations corresponding to stronger predictions. Predictions were checked against the MIPS database (see Methods), and the results are shown in Figure 3. This figure shows, for each of the methods, the percentage of predictions validated by MIPS as a function of the number of predictions made (when ordered from strongest to weakest correlation). In general, the methods exhibited the highest validation for their strongest predictions. For up to 600 predicted relationships (four per gene, on average), NMF far outperformed all other methods. For instance, for the 100 strongest predictions, the reliability in the NMF space was ~35%, whereas for all other spaces, including the original gene expression space, only 15%–25% of the predictions were validated. Beyond 800 predicted relationships, correlations in the original space did almost as well as NMF. However, the false positive rate at this level of prediction is likely to be too high to be useful. The reliability of predictions dropped off sharply for all spaces and eventually reached 9%, which was the probability of making a true prediction from the data set by chance. It might be assumed that some of the improvement in prediction score would be due to the sparsification procedure. However, it is shown that NMF without the sparsification procedure still outperforms the

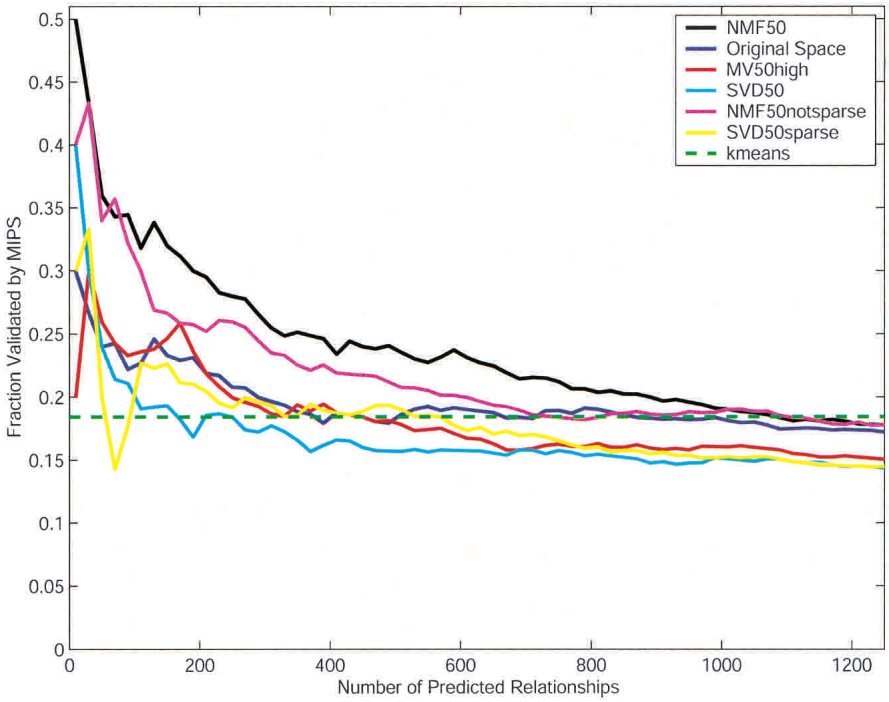

Figure 3 Performance of different spaces at predicting functional relationships between experiments with comparison to the MIPS classification of the deleted genes. (NMF50) NMF space with 50 basis vectors; (Original Space) original gene expression space; (SVD50) SVD space with 50 eigenvectors; (MV50high) space of the 50 most varying genes; (NMF50notsparse) NMF space with 50 basis vector without the sparsification procedure; (SVD50sparse) SVD sparsified; (k-means) predictions taken from k-means clustering with 50 clusters (3176 relationships).

other methods, and that a sparse SVD procedure performs similarly to SVD without sparsification.

A second and independent method was used to evaluate the predictions of functional relationships produced by NMF by comparing with data compiled in the Yeast Proteome Database (YPD; Costanzo et al. 2001). For the purposes of this study, the relationships reported by YPD were categorized as hard (indicating a direct measure of interaction, such as binding, or participation in the same pathway) and soft (indicating an indirect detection, such as coexpression). Examination of the strongest 100 predictions from NMF found that 58% were validated by querying YPD (38% were hard and 20% were soft functional relationships). This compared with ~35% of the same set that were verified through the MIPS database. The 58 validated predictions are listed in Table 3, and the 42 predictions that were not validated (but are testable predictions nonetheless) are listed in Table 4. Applying the same procedure to the strongest 100 predictions from the original gene expression space produced only 31% that could be verified by YPD (19% were hard and 12% soft). Thus, using the Yeast Proteome Database, dimensionality reduction through NMF appeared to be roughly twice as productive in predicting functional relationships as correlation in the original space of the data.

Table 3. The 58 Predictions That Could Be Validated by YPD of the 100 Strongest Functional Relationships Detected by NMF

Coregulated	
dfr1	ecm34
gyp1	yap7
ade16	sir1
hpt1	sir1
rml2	ymr293c
cbp2	mrpl33
mrpl33	rml2
cnb1	yor072w
ade16	ymr041c
gfd1	utr4
cla4 (haploid)	KAR2 (tet promoter)
yel001c	ymr141c
ckb2	gcn4
arg5,6	rpl8a
mrt4	rpl12a
clb6	whi2
erp2	ymr141c
erp2	yel001c
erp2	yor015w
rpl12a	yel033w
ckb2	rtg1
eca39	ras1
Identical Genes	
isw1	isw1, isw2
dig1, dig2	dig1, dig2 (haploid)
fks1 (haploid)	FKS1 (tet promoter)
bub3	bub3 (haploid)
Binding	
cla4 (haploid)	CDC42 (tet promoter)
qcr2 (haploid)	rip1
far1 (haploid)	ste4 (haploid)
bub1 (haploid)	bub3
bub1 (haploid)	bub3 (haploid)
Cell Wall	
fks1 (haploid)	2-deoxy-D-glucose
2-deoxy-D-glucose	Glucosamine
gas1	Tunicamycin
fks1 (haploid)	Glucosamine
yer083c	Tunicamycin
ste12 (haploid)	ste5 (haploid)
Mating	
ste5 (haploid)	ste7 (haploid)
fus3, kss1 (haploid)	ste5 (haploid)
ste18 (haploid)	ste5 (haploid)
ste12 (haploid)	ste18 (haploid)
ste18 (haploid)	ste7 (haploid)
fus3, kss1 (haploid)	ste18 (haploid)
fus3, kss1 (haploid)	ste7 (haploid)
fus3, kss1 (haploid)	ste12 (haploid)
ste12 (haploid)	ste7 (haploid)

(continued)

DISCUSSION

Here, NMF, a new machine-learning approach capable of identifying localized features in complex data sets, was applied to the analysis of microarray data from a series of 300 yeast experiments (of which 276 were deletion strains; Hughes et al. 2000). The essence of NMF is that the algorithm must choose a small number of features (basis vectors) to act as building blocks that can be scaled and added together in various combinations to best reconstruct the original data. Restriction to a small number of basis vectors causes the algorithm to select patterns of genes that occur frequently in the data. The application of a data analysis approach that extracts localized data features from a set of experiments that span a wide range of genetic variation holds the potential to be a particularly powerful method to detect functional cellular subsystems (the features encoded in the basis vectors) as well as individual pairwise functional genetic relationships.

The experimental variation sampled by the 300 experiments could be well represented with just 50 features. Moreover, this set of 50 features encoded in the basis vectors tended to correspond to sets of known functional genetic groupings of genes. Large numbers of genes involved in similar or related cell functions appeared together due to a local similarity in their expression profiles. It should be noted that because of the limited data (i.e., not all yeast deletion strains were sampled), not all cellular functions were identified. Some cellular systems were sampled more in the experiments than others. For example, the mating and pheromone grouping is particularly well identified. Basis vector 8 consisted mostly of genes involved in mating and even contained six verified targets of *STE12* that were not identified by previous studies.

Conventional clustering techniques focus on elucidating groupwise relationships among genes by sorting them according to a pairwise similarity metric. NMF procedures applied here also identify groups of genes related to one another in expression patterns and form them together into basis vectors. It is clear that genes in the same cluster have similar expression patterns. In the case of NMF basis vectors, the relationship is less clear; treating the contribution of a set of basis vector genes as a group is an efficient representation of the expression data. This may or may not be a good indicator of biological relevance.

Table 3. *Continued*

Ergosterol Pathway	
erg3 (haploid)	Itraconazole
erg2	Itraconazole
yer044c (haploid)	ERG11 (tet promoter)
ERG11 (tet promoter)	Itraconazole
erg3 (haploid)	ERG11 (tet promoter)
erg3 (haploid)	yer044c (haploid)
erg2	erg3 (haploid)
erg2	yer044c (haploid)
erg2	ERG11 (tet promoter)
Vacuolar ATPase	
cup5	mac1
mac1	vma8
cup5	vma8

Coregulated genes were found to be coregulated by other functional genomics studies. Binding refers to genes whose proteins have been shown to bind each other. Cell Wall, Mating, and Ergosterol Pathway are all genes that have been experimentally shown to be involved in the named cellular function.

Pairwise relationships between experiments were evaluated by locating pairs of experiments that were constructed from the same NMF building blocks (basis vectors). With this detection scheme, NMF was superior to any other method examined, including other sets of 50 basis vectors constructed from other procedures, as well as standard correlations in the full gene expression space of the original experiments. The initial analysis of this same compendium data set reported by Hughes et al. (2000) used conventional clustering methods and found a series of very interesting and useful relationships between genes. Many of the genes that were clustered together in that study were also scored as related in the current work. For example, the sections headed Ergosterol Pathway, Cell Wall, Mating, and Vacuolar ATPase in Table 3 contain many relationships also detected by Hughes et al. (2000) by use of more standard techniques; however, most of the other validated relationships evaded detection by conventional techniques. Specifically, Hughes et al. (2000) found 38 of the 58 relationships. This includes the section headed Binding in Table 3, for which particularly strong experimental validation is available.

Figure 4 illustrates the increased similarity seen in the NMF feature space compared with that seen in the original data space for four pairwise functional relationships from Tables 3 and 4. Two of these (yer084w:*SBH2* and ymr025w:ymr029c) were not corroborated by YPD, whereas the other two (*STE5:STE11* and *RTS1:RTG1*) are each known to be functional relationships. As the numerical values in the figure indicate, the correlation in NMF space was significantly higher than in the original gene expression space. Essentially, this stems from the fact that NMF recognized the expression patterns of strains deleted for the genes in question as being constructed of very similar sets of building blocks, and the correlation in the expression pattern was larger for the genes comprising these building blocks. For instance, the expression profiles for strains deleted in *STE11* and in *STE5* were each dominated by basis vector 8 (the building block consisting largely of mating genes) and had relatively small (but still correlated contributions) from other basis vectors. NMF recognized this local similarity across some genes, whereas most clustering algorithms would focus only on the global similarity of the expression profile. Comparing the same two strains in the original data space shows that their gene expression patterns were highly correlated for some genes but not for others. Therefore, NMF is a way to focus on the functionally important parts of gene expression profiles.

Note that due to the fact that all values in NMF space are non-negative by definition, the distribution of correlations is somewhat shifted toward higher values and it has a longer tail (see Supplemental Material Fig. S-1, available online at www.genome.org). However, this effect alone does not explain the higher correlation coefficients found in NMF versus in the original space, seen in Figure 4. The correlations found in the NMF space occur at a higher percentile than those in the original space. For example, a correlation of 0.8 corresponded to a 99.90 (99.93) percentile in the NMF (original) space; a correlation of 0.4 corresponded to a 95.00 (98.00) percentile in the NMF (original) space. The differences in correlation coefficients observed here correspond to values in the neigh-

Table 4. The 42 Predictions of Functional Relationships That Could Not Be Verified on YPD From the 100 Strongest Relationships Detected

rtg1	vps8
are1, are2 (haploid)	yor015w
pex12	yea4
ckb2	yel008w
yer002w	ymr034c
mrt4	yel033w
ckb2	rts1
mrpl33	ymr293c
imp2	yer050c
cbp2	pet111
cyt1	pet111
yer034w	ynd1
rps24a	ymr014w
yel001c	yor015w
ymr014w	yor006c
aep2	rml2
aep2	mrpl33
ymr014w	yor078w
rml2	yer050c
mrpl33	yer050c
aep2	imp2
sir1	ymr041c
ymr034c	yor015w
pfd2	yor051c
ymr025w	ymr029c
ckb2	vps8
msul	ymr293c
sbh2	yer084w
mrpl33	msu1
imp2	ymr293c
rtg1	rts1
msu1	yer050c
msu1	rml2
yml003w	ymr034c
aep2	msu1
CDC42 (tet promoter)	KAR2 (tet promoter)
rps24a	yor078w
pfd2	yel044w
gcn4	yel008w
yer050c	ymr293c
aep2	yer050c
aep2	ymr293c

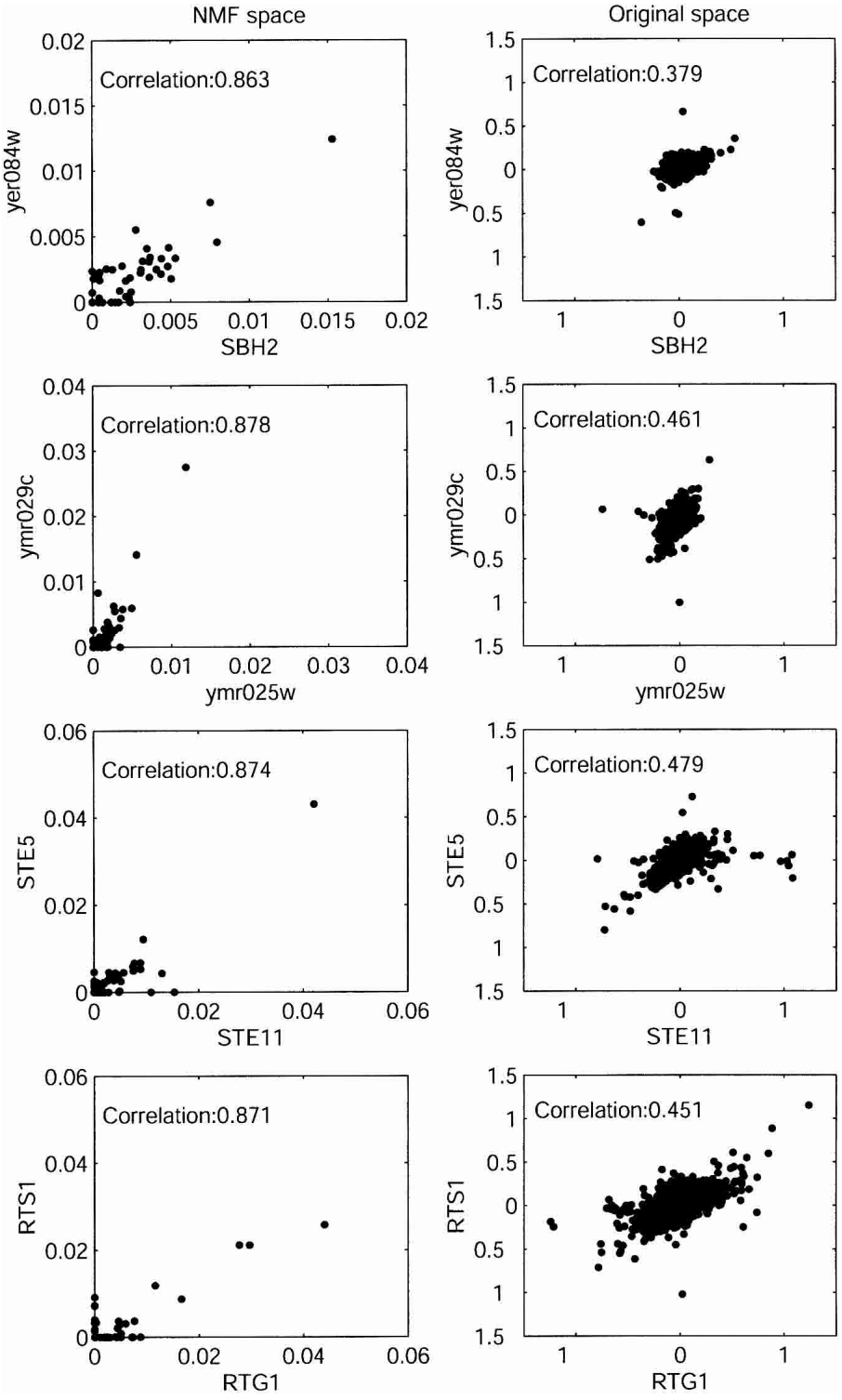

Figure 4 Correlation for four illustrative pairwise functional genetic relationships. For comparison, the correlation plot of the pair of experiments in NMF space is shown at *left*, and in the original gene space at *right*.

borhood of 0.8 (99.90 percentile) in NMF space and only 0.4 (98.00 percentile) in the original space. Thus, even though the distribution is somewhat shifted in the NMF space, the higher correlation coefficients observed do correspond to higher significance, as indicated by percentile ranking.

In Table 4 are listed 42 predictions of functional relationships detected by NMF, but not present in YPD. Some predicted relationships are between genes classified as mitochondrial (e.g., *AEP2:YER050C* and *MSU1:MRPL33*), just as some of the verified relationships are between mitochondrial genes (e.g., *RML2:YMR293C*). Moreover, a number of small networks of mitochondrial genes occur in the strongest 100 NMF relationships; most genes in these networks were clustered together in the original analysis of the data by Hughes et al. (2000). Another tight network of functional relationships can be seen among *CKB2*, *YEL008W*, *GCN4*, *RTS1*, *RTG1*, and *VPS8*, some of which were and some of which were not verified by YPD. The existence of these tight interconnected relationships adds to the likelihood that the predictions are correct.

While this manuscript was in preperation, two studies (Gavin et al. 2002; Ho et al. 2002) that focused on the large-scale identification of protein–protein interactions in yeast were published. For the study by Ho et al. (2002) data was readily available online, and we used it as an additional means of verifying predictions. There was little overlap in scope with our study, as of a total of identified 8114 interactions, only 74 fall within the set of 276 gene deletions in our study. Allowing for one connecting link between interacting proteins, this number increased to 2001. Of our best 100 predictions, an additional two were verified (between *CKB2* and *VPS8* and between *CDC42* and *KAR2*). It should be noted that both come from the set of 42 interactions in which no direct link was found on YPD (i.e., none of the 58 predictions that were verified on YPD were found by Ho et al. 2002). The little overlap with our predictions may thus stem from the sparseness of both data sets. When scoring our predictions using the interactions found by Ho et al. (2002), we still find that predictions made using NMF have a higher likelihood of being correct than ones made from pure correlation (data not shown).

One feature of the approach taken here is that pairwise relationships were only scored for genes that had been directly manipulated in the experiments (deleted or overexpressed). As described in the Methods section, NMF can also be applied to detect relationships between genes that have been monitored using expression arrays, but not directly manipulated experimentally. Preliminary studies using NMF in this mode suggest that it is again superior in detecting functional genetic relationships compared with approaches that apply clustering or correlation directly in the original data space. A further shortcoming that remains, however, is the elimination from analysis of genes whose expression is undetectable in the control experiments (to avoid division by zero). Functional relationships involving such genes (comprising roughly one-sixth of the genome for the current data set) cannot be scored. In future studies, it may be possible to insert a minimal expression level for such genes in the control experiment, although further work is necessary to see whether this introduces other problems, such as feature misscaling.

In the current study, no separate attempt was made to smooth or filter the data set to reduce or eliminate the effects of experimental noise or error. In some sense, NMF itself performs a smoothing function on the data through factorization and reconstruction. Features that appear consistently in the data set are selected out to become basis vectors, whereas features that appear inconsistently in the data due to experimental variability or other factors tend to be smoothed. For the results reported here, only genes with no detectable expression in the control experiment were removed. When more stringent significance filters were applied to the data, the results remained similar (data not shown).

Of the 50 basis vectors resulting from this analysis, many were sparse (that is, they represented features consisting of a relatively small number of genes). However, some basis vectors were not sparse and contained too many genes to be easily annotated as associated with a small number of cellular functions. The NMF algorithm could be modified to enforce sparser basis vectors; alternatively, it is anticipated that larger data sets will result in basis vectors that are more uniformly sparse and may correspond to smaller features. An advantage of NMF is that it is expected to be a better detector of features when confronted with larger data sets.

METHODS

General Approach

Data from a set of expression array experiments were represented as a single matrix $\overleftrightarrow{V}$. Each column corresponded to the processed intensities from one experiment; each element of a column was derived from the intensity for one gene probe in the corresponding experiment. A row of the matrix corresponded to the processed intensity for a single gene probe across all experiments. An $n \times m$ matrix $\overleftrightarrow{V}$ corresponded to m arrays (i.e., experiments) in which measurements were made for the same n genes in each. The major analysis method applied here, NMF, corresponded to an approximate factorization of the matrix $\overleftrightarrow{V}$ into a pair of matrices $\overleftrightarrow{W}$ and $\overleftrightarrow{H}$.

$$\overleftrightarrow{V} \approx \overleftrightarrow{W} \cdot \overleftrightarrow{H} \tag{1}$$

The factorization was chosen with a particular rank, k, so that $\overleftrightarrow{W}$ was of dimension $n \times k$ and $\overleftrightarrow{H}$ was $k \times m$. In the work described here, k was chosen to be relatively small compared with the dimensions of the original data $\overleftrightarrow{V}$ (i.e., $k \cdot (m + n) < n \cdot m$), so the factorization was approximate and corresponded to a compression of the data. Moreover, the factorization could be viewed as a representation of the data in a new space of lower dimensionality (k). There are two equally valid interpretations of the dimensionality reduction. One is that the columns of $\overleftrightarrow{W}$ were "basis experiments" (having the dimensionality of a single array or experiment) and each row of $\overleftrightarrow{H}$ was the representation of a particular experiment in the new k-dimensional space. Alternatively, the rows of $\overleftrightarrow{H}$ were "basis genes," and each column of $\overleftrightarrow{W}$ then corresponded to a representation of a particular gene in the new space. The unique feature of NMF is that none of the matrices in equation 1 ($\overleftrightarrow{V}$, $\overleftrightarrow{W}$, or $\overleftrightarrow{H}$) are permitted to have negative entries (Lee and Seung 1999).

Implementation of NMF

The NMF algorithm was coded using the mathematics and matrix algebra package MATLAB version 6 (R12) (Mathworks, Inc.). The key features of the algorithm involved iteratively improving matrices $\overleftrightarrow{W}$ and $\overleftrightarrow{H}$ to improve the approximation to $\overleftrightarrow{V}$ while maintaining non-negative matrix entries throughout. This was achieved using an update-rule approach (Lee and Seung 1999). For a given value of the NMF dimensionality k, the algorithm was started with random matrices $\overleftrightarrow{W}$ and $\overleftrightarrow{H}$. The random initial seed was a uniform distribution of real numbers from 0 to 1 for all matrix elements of $\overleftrightarrow{W}$ and $\overleftrightarrow{H}$. The two matrices were iteratively updated using the rules,

$$\overleftrightarrow{H}_{a\mu} \leftarrow \overleftrightarrow{H}_{a\mu} \frac{(\overleftrightarrow{W}^T \overleftrightarrow{V})_{a\mu}}{(\overleftrightarrow{W}^T \overleftrightarrow{W} \overleftrightarrow{H})_{a\mu}} \tag{2}$$

$$\overleftrightarrow{W}_{ia} \leftarrow \overleftrightarrow{W}_{ia} \frac{(\overleftrightarrow{V} \overleftrightarrow{H}^T)_{ia}}{(\overleftrightarrow{W} \overleftrightarrow{H} \overleftrightarrow{H}^T)_{ia}} \tag{3}$$

which minimize the root-mean-square (RMS) error ($E = \| \overleftrightarrow{V} - \overleftrightarrow{W} \cdot \overleftrightarrow{H} \|_2$) between the actual data $\overleftrightarrow{V}$ and the reduced-dimension reconstruction of the data ($\overleftrightarrow{W} \cdot \overleftrightarrow{H}$; Lee and Seung 2001). Because the update rules were multiplicative, initial non-negative matrices remained non-negative for all future iterations. Iterations were continued until the RMS error change in an iteration was <0.1 in absolute RMS error, which corresponded to roughly 0.005% of the final RMS error.

The update rules corresponded to a form of gradient descent, and thus, found only a local minimum. To address this limitation, the procedure was repeated 100 times, starting with different initial matrices. The factorization leading to the lowest RMS error was used in further analysis. Studies were carried out for values of the NMF dimensionality (k) ranging from 10–80. The solutions found were reproducible; basis vectors from factorizations that differed in the initial matrices showed correlation coefficients of >0.90.

A single NMF factorization for a 5346 × 300 data set required ~30 min of CPU time on a 500 MHz Pentium III workstation and occupied roughly 70 MB of memory. The current implementation was dominated by matrix multiplication, leading to computation times that scaled as the number of matrix entries raised to roughly the power 1.35 (typical of matrix multiplication in MATLAB and other modern packages). The relative simplicity of the update-rule implementation does not require first- or second-derivative information, which would add significantly to memory usage. Memory requirements scaled linearly with data set size due to the need to store data and factor matrices.

Trial implementations on smaller test problems were also carried out with nonlinear optimizers CONOPT2 version 2.071G (ARKI Consulting & Development A/S) and LOQO

version 4.01 (Princeton University); the values of matrix elements in W and H were optimized directly and subject to non-negativity constraints to minimize the RMS error. Although quite successful on small problems, these methods require additional memory for storage of the gradient, and were thus not feasible for the data set analyzed here.

To ensure sparsity of the resulting basis vectors, the most significant genes for every basis experiment were selected so as to produce an average of 5% of the entries used across all basis vectors. Operationally, this was achieved by allowing a fixed percentage (9.7%) of the maximum gene to be significant (non-zero). After selecting of the most significant genes, all other genes were constrained to zero, and the resulting sparsified basis vectors were reoptimized to convergence using the update rules in equations 2 and 3. This sparsification procedure was found to be the best performing one from many procedures tried, when minimizing both the overall RMSD of the factorization and the number of signifcant genes for each basis vector. The RMSD of the sparsified factorization was 2132 versus 1702 for the factorization without sparsification (a difference of 25%), whereas only 5% of the entries are used as significant genes. As can be seen in Figure 3, the sparsification has a minor effect on the performance of the algorithm.

In separate calculations, SVD of data matrices was carried out using the built-in functionality in MATLAB. A representation of an SVD factorization of rank k corresponded to using only the k highest eigenvalues. Absolute RMS error values were calculated for the same data set and the same ranks as for NMF. Furthermore, as a control, SVD was carried out on random matrices composed of vectors of the same mean and standard deviation as the sample data.

Annotating Basis Vectors

The functional categorizations available at the Munich Information Center for Protein Sequences (the MIPS categories) were used to assign genes to biochemical pathways or cellular function. There are a total of 107 MIPS categories that cover different metabolic pathways, such as the TCA cycle and glycolysis, as well as different cellular functions, such as cell membrane biosynthesis or mating (Mewes et al. 2000). Some of the categories overlap (for example, the category glycolysis is a subset of the category energy metabolism), and one gene can be assigned to more than one category.

Each basis vector (basis experiment) was annotated with the MIPS categories that dominated its makeup by comparing the frequency with which genes from each category appeared in a basis vector with that expected from a random distribution. One million genes were selected at random from the same set of genes present in the experimental data. The corresponding MIPS categories were identified, and the mean and the standard deviation of occurrence was calculated for every category. This procedure was carried out twice to ensure convergence of the random distribution. If the occurrence of a particular MIPS category in a basis vector exceeded the mean of the random occurrence by more than five times its standard deviation (a 5σ cutoff), this particular category was assigned to the basis vector as enriched. As a negative control, basis vectors were generated from random numbers and subjected to the same significance cutoffs and annotation procedure. In 1000 random basis vectors, no category was ever assigned as being enriched.

Predicting Functional Relationships

An important test of the utility of data reduction using nonnegative matrix factorization was to assess its ability to predict functional relationships between genes. To predict functional relationships between genes or experiments on the basis of expression data, it is typical to assume that similarity in expression suggests a functional relationship between genes or experiments. Here, the same assumption was made both in the original space of the data and in the reduced dimensional spaces, such as that computed by NMF. The Pearson correlation coefficient was calculated between all genes (or all experiments), and the absolute value of the correlation coefficient was used as a predicted score for the relationship. This method scored positive and negative correlations equally strongly.

In the data set used, most experiments corresponded to deletion mutants of a specific gene, so that functional relationships between experiments in turn implied functional relationships of the deleted genes. Other experiments corresponded to the overexpression of genes, which again linked the experiments directly to the gene in question. The rest of the experiments corresponded to treatment with a well-characterized drug. Those experiments then linked the response in expression pattern to the functional mechanism of this particular drug.

To judge the predicted functional relationships between genes required some set of true relationships. For this purpose, existing bioinformatic databases were used, although clearly such data are largely incomplete and may not be fully verified. The two databases used were the MIPS categorization (Mewes et al. 2000) and the YPD (Costanzo et al. 2001). Two genes appearing in the same MIPS category were scored as functionally related (e.g., two genes encoding ribosomal proteins). The MIPS categorization was checked for every gene in the data set, as well as for every gene for which there was a deletion or overexpressed mutant in the data set, yielding a list of validated interactions. Predictions from the correlation score in NMF space were compared with this list, starting with the predictions of highest correlation.

The functional relationships predicted from gene expression data using NMF were compared with functional relationships predicted from other approaches. The same analysis and validation procedure was applied to the correlation score in five other spaces: the original full-dimensionality of the experimental space, reduced dimensionality using SVD with the 50 most significant dimensions, reduced dimensionality using only the 50 most variable genes in the data set, reduced dimensionality using 50 NMF basis vectors that were not sparsified, and reduced dimensionality using 50 eigenvectors from SVD that were sparsified. The value of 50 was chosen to compare different same-sized reduced-dimension representations of the data to that from NMF.

Eigenvectors from SVD were sparsified using the above procedure. The encodings were then obtained using the pseudoinverse. As a further comparison, k-means clustering was carried out using the euclidian distance metric starting from a random initial seed of cluster centers and interatively updating the center positions.

A second and independent method of scoring predicted functional relationships used the Yeast Proteome Database (YPD; Costanzo et al. 2001). YPD contains detailed information about genetic or physical interaction, functional relationships, and coregulation of all genes in yeast. The information in YPD is based on a large number of papers from the scientific literature. Results cataloged in YPD include those from biophysical, molecular biological, genetic, and functional genomic experiments. The strongest 100 predictions from NMF and from correlations in the original experimental data space were scored against YPD. Any link in YPD between two genes (e.g., coregulation, genetic interaction, or binding) was viewed as a validation of the prediction. Moreover, for the soft validations, one linking gene was permitted. That is, if gene A interacted with gene Z, and gene B was coregulated with gene Z in YPD, then gene A and B were scored as coregulated. For this purpose, at least one of the two relations was required to be a hard interaction.

Data Source and Preprocessing

This study is based on the analysis of a large, publicly available microarray data set from Rosetta Inpharmatics, Inc. encompassing genome-wide expression data of *S. cerevisiae* in 276 deletion mutants, 11 tetracycline regulated alleles of essential genes (overexpression) and 13 wild-type strains treated with well-characterized drugs (a total of 300 experiments; Hughes et al. 2000). All experiments used the Saccharomyces Genome Deletions Consortium strain background. Most of the deletion mutants were diploid mutants (i.e., both alleles were deleted from the genome). For some essential genes, haploid mutants were made. This impaired, but did not remove, the gene function. The strains were grown according to a standard protocol and in parallel with corresponding wild-type control cultures.

Gene expression was measured using spotted microarrays, giving the ratio of expression in the mutant (or drug-treated) strain relative to the gene expression in the control (wild-type) experiments. The spotted arrays measured expression for a total of 6316 ORFs; the data set was 6316 genes by 300 experiments (data available from Rosetta Inpharmatics Inc. at http://www.rii.com/register/cell2000102Hughes/EULA.htm). It is likely that much of the yeast gene expression space is sampled in this data set, which spans very different conditions; thus, it appears a good data source in which to seek gene expression features.

The log-transformed ratios were used as input data for our algorithm; the transformed ratios ranged from approximately -3 (1000 times down-regulated with respect to the control experiment) to +3 (1000 times up-regulated). Some genes had no detectable expression in the control experiment and were removed from further analysis to prevent division by zero. The resulting data set contained 5346 genes. To make the data fit the constraint of non-negativity, the data were folded. Every gene was represented in two rows of the matrix, the first occurrence to indicate positive expression relative to wild type, and the second to indicate negative. This effectively doubled the size of the data set (to 10692 genes). In any one experiment, the log-expression ratio for every gene was either positive (i.e., the gene was up-regulated with respect to the control experiment) or negative. The resulting data matrix was of size 10692×300, and half of its entries were equal to zero. This procedure was necessary, as NMF performs most optimally on sparse data sets. A simple shifting procedure, that is, adding a fixed constant to each matrix element to make all positive, would create a positive, but very non-sparse matrix, and hence, was inappropriate. For reconstructing the data, we simply reversed this procedure by subtracting the row corresponding to down-regulation from the row corresponding to up-regulation. Correlations were computed by operating on vectors of length 10692, with no special treatment for paired entries involving the up- and down-regulation of the same gene. Interestingly, in each basis vector, the same gene was never represented as both up- and down-regulated.

ACKNOWLEDGMENTS

We thank H. Sebastian Seung, Michael D. Altman, Justin A. Caravella, Gerald R. Fink, David F. Green, Chris Kaiser, Sriram Kosuri, Douglas A. Lauffenburger, Robert T. Sauer, Anthony J. Sinskey, Peter K. Sorger, and Shari Spector for helpful discussions and suggestions. We also thank the two anonymous referees for insightful comments. This work was partially supported by the Alfred P. Sloan Foundation and the National Institutes of Health (MH62344). P.M.K. was supported by a Merck/MIT Graduate Fellowship and a Ph.D. Fellowship from the Boehringer Ingelheim Fonds.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., and Levine, A.J. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci.* **96:** 6745–6750.

Alter, O., Brown, P.O., and Botstein, D. 2000. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci.* **97:** 10101–10106.

Bittner, M., Meltzer, P., Chen, Y., Jiang, J., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., et al. 2000. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406:** 536–540.

Broet, P., Richardson, S., and Radvanyi, F. 2002. Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *J. Comput. Biol.* **9:** 671–683.

Brown, C.S., Goodwin, P.C., and Sorger, P.K. 2001. Image metrics in the statistical analysis of DNA microarray data. *Proc. Natl. Acad. Sci.* **98:** 8944–8949.

Brown, M.P.S., Grundy, W.N., Lin, D., Christiani, N., Sugnet, C.W., Furey, T.S., Ayres Jr., M., and Haussler, D. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci.* **97:** 262–267.

Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., et al. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2:** 65–73.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O., and Herskowitz, I. 1998. The transciptional program of sporulation in budding yeast. *Science* **282:** 699–705.

Coller, H.A., Grandori, C., Tamayo, P., Colbert, T., Lander, E.S., Eisenman, R.N., and Golub, T.R. 2000. Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. *Proc. Natl. Acad. Sci.* **97:** 3260–3265.

Costanzo, M.C., Crawford, M.E., Hirschman, J.E., Kranz, J.E., Olsen, P., Robertson, L.S., Skrzypek, M.S., Braun, B.R., Hopkins, K.L., Kondu, P., et al. 2001. YPD, PombePD, and WormPD: Model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res.* **29:** 75–79.

DeRisi, J.L., Iyer, V.R., and Brown, P.0. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278:** 680–686.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95:** 14863–14868.

Ferea, T.L., Botstein, D., Brown, P.O., and Rosenzweig, R.F. 1999. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc. Natl. Acad. Sci.* **96:** 9721–9726.

Fodor, S.P.A., Rava, R.P., Huang, X.H.C., Pease, A.C., Holmes, C.P., and Adams, C.L. 1993. Multiplexed biochemical assays with biological chips. *Nature* **364:** 555–556.

Gasch, A.P. and Eisen, M.B. 2002. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.* **3:** 1–22.

Gavin, A.C., Boesche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415:** 141–147.

Getz, G., Levine, E., and Domany, E. 2000. Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci.* **97:** 12079–12084.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286:** 531–537.

Granjeaud, S., Bertucci, F., and Jordan, B.R. 1999. Expression profiling: DNA arrays in many guises. *BioEssays* **21:** 781–790.

Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., and Young, R.A. 2001. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.* **6:** 422–433.

Heyer, L.J., Kruglyak, S., and Yooseph, S. 1999. Exploring expression data: Identification and analysis of coexpressed genes. *Genome*

Res. **9:** 1106–1115.
Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., et al. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415:** 180–183.
Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., and Young, R.A. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95:** 717–728.
Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennet, H.A., Coffey, E., Dai, H., He, Y.D., et al. 2000. Functional discovery via a compendium of expression profiles. *Cell* **102:** 109–126.
Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Baumgarner, R., Goodlett, D.R., Aebersold, R., and Hood, L. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292:** 929–934.
Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Hudson Jr., J., Boguski, M.S., et al. 1999. The transcriptional program in the response of human fibroblasts to serum. *Science* **283:** 83–87.
Kim, S.K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J.M., Eizinger, A., Wylie, B.N., and Davidson, G.S. 2001. A gene expression map for *Caenorhabditis elegans*. *Science* **293:** 2087–2092.
Lee, D.D. and Seung, H.S. 1999. Leaming the parts of objects by non-negative matrix factorization. *Nature* **401:** 788–791.
———. 2001. Algorithms for non-negative matrix factorization. *Adv. Neural Info. Proc. Syst.* **13:** 556–562.
Li, C. and Wong, W.H. 2001. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci.* **98:** 31–36.
Mewes, H.W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schuller, C., et al. 2000. MIPS: A database for genomes and protein sequences. *Nucleic Acids Res*. **28:** 37–40.
Misra, J., Schmitt, W., Hwang, D., Hsiao, L.L., Gullans, S., Stephanopoulos, G., and Stephanopoulos, G. 2002. Interactive exploration of microarray gene expression patterns in a reduced dimensional space. *Genome Res*. **12:** 1112–1120.
Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290:** 2306–2309.
Schena, M., Schalon, D., Davis, R.W., and Brown, P.0. 1995. Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. *Science* **270:** 467–470.
Sherlock, G. 2000. Analysis of large-scale expression data. *Curr. Opin. Immun*. **12:** 201–205.
Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9:** 3273–3297.
Tamayo, P., Slomin, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., and Golub, T.R. 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci*. **96:** 2907–2912.
Tanay, A. and Shamir, R. 2001. Computational expansion of genetic networks. *Bioinformatics* **17:** S270–S278.
Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, G.M. 1999. Systematic determination of genetic network architecture. *Nat. Genet.* **22:** 281–285.
Zhu, Z., Pilpel, Y., and Church, G.M. 2002. Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm. *J. Mol. Biol*. **318:** 71–81.

WEB SITE REFERENCES

http://mips.gsf.de; Munich Information Center for Protein Sequences.
http://www.incyte.com/; Yeast Proteome Database.
http://www.rii.com/register/cell2000102Hughes/EULA.htm; Source Data at Rosetta Inpharmatics.

Received October 11, 2002; accepted in revised form March 24, 2003.

Subsystem Identification Through Dimensionality Reduction of Large-Scale Gene Expression Data

Philip M. Kim and Bruce Tidor

Genome Res. 2003 13: 1706-1718
Access the most recent version at doi:10.1101/gr.903503