Research

From co-expression to co-regulation: how many microarray experiments do we need?

Ka Yee Yeung*, Mario Medvedovic† and Roger E Bumgarner*

Addresses: *Department of Microbiology, University of Washington, Seattle, WA 98195, USA. †Center for Genome Information, Department of Environmental Health, University of Cincinnati Medical Center, Cincinnati, OH 45267, USA.

Correspondence: Ka Yee Yeung. E-mail: kayee@u.washington.edu. Roger E Bumgarner. E-mail: rogerb@u.washington.edu

Published: 28 June 2004

Genome ***Biology*** 2004, **5:**R48

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/7/R48

Received: 18 February 2004
Revised: 19 April 2004
Accepted: 28 May 2004

Abstract

Background: Cluster analysis is often used to infer regulatory modules or biological function by associating unknown genes with other genes that have similar expression patterns and known regulatory elements or functions. However, clustering results may not have any biological relevance.

Results: We applied various clustering algorithms to microarray datasets with different sizes, and we evaluated the clustering results by determining the fraction of gene pairs from the same clusters that share at least one known common transcription factor. We used both yeast transcription factor databases (SCPD, YPD) and chromatin immunoprecipitation (ChIP) data to evaluate our clustering results. We showed that the ability to identify co-regulated genes from clustering results is strongly dependent on the number of microarray experiments used in cluster analysis and the accuracy of these associations plateaus at between 50 and 100 experiments on yeast data. Moreover, the model-based clustering algorithm MCLUST consistently outperforms more traditional methods in accurately assigning co-regulated genes to the same clusters on standardized data.

Conclusions: Our results are consistent with respect to independent evaluation criteria that strengthen our confidence in our results. However, when one compares ChIP data to YPD, the false-negative rate is approximately 80% using the recommended *p*-value of 0.001. In addition, we showed that even with large numbers of experiments, the false-positive rate may exceed the true-positive rate. In particular, even when all experiments are included, the best results produce clusters with only a 28% true-positive rate using known gene transcription factor interactions.

Background

Cluster analysis is a popular exploratory technique to analyze microarray data. It is often used for pattern discovery - to identify groups (or clusters) of genes or experiments with similar expression patterns. Cluster analysis is an unsupervised learning approach in which genes or experiments are assigned to groups (or clusters) based on their expression patterns and no prior knowledge of the data is required. A common application of cluster analysis is to identify potentially meaningful relationships between genes or experiments or both [1-3].

Transcription of a gene is determined by the interaction of regulatory proteins (that is, transcription factors) with DNA sequences in the gene's promoter region [4]. A common application of cluster analysis is to identify potential transcriptional modules, for example genes that share common promoter sites. An example of this is the large-scale analysis of gene expression as a function of cell cycle in yeast [5]. The study focused on genes that behaved similarly to other genes that are known to be regulated during the cell cycle. A total of 800 genes were found to be regulated during the cell cycle, and 700 base pairs (bp) of genomic sequence immediately upstream of the start codon for each of these 800 genes was analyzed to identify potential binding sites for known or novel factors that might control expression during the cell cycle. The majority of the genes were shown to have good matches to known cell-cycle transcription factor binding sites.

The approach pioneered by Spellman *et al.* [5] - for example the meta-analysis of massive amounts of gene-expression data to identify genes that are co-expressed followed by promoter analysis - is now commonplace [6-10]. Cluster analysis is often used to identify genes whose expression levels are correlated across numerous experiments. However, using cluster analysis to infer regulatory modules or biological function has its limitations. In general, cluster analysis always returns clusters independent of the biological relevance of the clusters. Microarray data can be quite noisy owing to measurement errors and technical variations, and cluster analysis will find patterns in noise as well as in signal. In this paper, we address two main questions. The first is how often do we discover co-regulated genes (that is, genes that are regulated by common transcription factors) from co-expressed genes (that is, genes that share similar expression patterns). The second asks how the following factors affect the likelihood of finding co-regulated genes: the number of microarray experiments in the microarray datasets; the clustering algorithm used; and the diversity of experiments in a microarray dataset.

The primary thrust of this paper is to provide guidance to researchers who wish to use cluster analysis of gene expression data to identify co-regulated genes. In particular, we provide an estimate of the accuracy of this association as a function of the number of experiments used in cluster analysis. This information is critical for researchers in assessing how much effort (if any) should go into promoter analysis of genes that cluster together in a fixed number of experiments.

Our approach

Our goal is to study the likelihood that co-expressed genes are regulated by the same transcription factor(s). We define co-expressed genes as genes that share similar expression patterns as discovered by cluster analysis, and we define co-regulated genes as genes that are regulated by at least one common known transcription factor. Our overall approach is illustrated in Figure 1. In brief, we first defined a set of genes that are controlled by known transcription factors. As there is

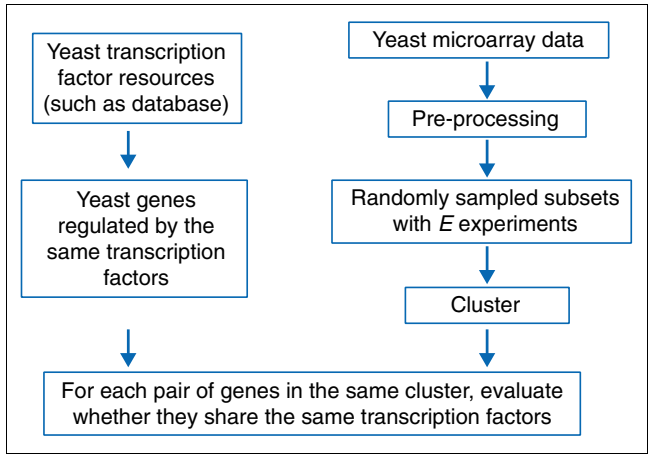

Figure 1
Our overall approach. We applied different clustering algorithms to cluster the genes in yeast microarray datasets with different sizes to identify co-expressed genes. The level of co-regulation is evaluated using yeast transcription factor databases (SCPD and YPD) and ChIP data. The clustering results are then evaluated by determining the fraction of gene pairs from the same clusters that share at least one known common transcription factor.

both an abundance of yeast array data and many available resources on yeast transcription factors such as yeast transcription factor databases [11,12] and yeast chromatin immunoprecipitation (ChIP) data [13,14], we ran our experiments on yeast data. Genes that share common transcription factors are taken as our 'gold standards' for evaluating the ability of cluster analysis to infer co-regulation. We then identified large publicly available yeast microarray datasets, preprocessed these datasets to remove genes with many missing values and created randomly sampled subsets of the data on which we performed cluster analysis. The randomly sampled subsets with different numbers of microarray experiments allow us to study the effect of the size of microarray datasets on the likelihood of discovering co-regulated genes. We used two publicly available yeast microarray datasets consisting of hundreds of microarray experiments: the yeast compendium data [2] and the yeast environmental stress data [15,16]. The yeast compendium dataset [2] consists of 300 knock-out microarray experiments, whereas the yeast environmental stress dataset [15,16] consists of 225 concatenated time course microarray experiments. We investigated the effect of different clustering algorithms on identifying co-regulated genes by applying different clustering algorithms to subsets of these microarray datasets, including heuristic-based clustering algorithms such as hierarchical complete-link and hierarchical average-link algorithms, and model-based clustering algorithms such as MCLUST [17-19] and the infinite mixture model-based method (IMM) [20-22].

We used two independent sources of data to define co-regulated genes: yeast transcription factor databases [11,12] and yeast chromatin immunoprecipitation (ChIP) data [13,14].

Transcription factor databases are based on published results in the literature and are generally based on specific measures of physical interactions between the transcription factor, promoter, and some measure that the transcription factor truly regulates the downstream gene. We used two different transcription factor databases: the *Saccharomyces cerevisiae* Promoter Database (SCPD) [11], and the Yeast Proteome Database (YPD) [12]. The SCPD lists approximately 230 yeast genes that are regulated by 90 transcription factors, while the YPD lists approximately 580 yeast genes that are regulated by 120 transcription factors, as of November 2001. We extracted two subsets of genes that are listed in the SCPD and YPD databases from each of the yeast compendium [2] and the environmental stress [15,16] microarray datasets. After eliminating genes and experiments with many missing values, the two gene subsets from the compendium data evaluated using SCPD and YPD consist of 215 genes under 273 experiments, and 537 genes under 258 experiments, respectively. The two gene subsets from the environmental stress data evaluated using SCPD and YPD consist of 205 genes under 205 experiments, and 526 genes under 198 experiments, respectively. The ChIP data represents a systematic technique to determine target genes bound to a set of transcription factors *in vivo*. However, the binding of a transcription factor to the promoter sequence of a gene does not necessarily imply that the transcription factor actually regulates the gene. We evaluated the two gene subsets from each of the yeast compendium and environmental stress datasets using the ChIP data [14] in addition to the corresponding transcription factor database. Two genes are considered co-regulated if they are bound to at least one common transcription factor in the ChIP data. The publicly available ChIP data [14] adopts an error model in which a confidence value (p-value) is assigned to each regulator-DNA interaction, and we used the recommended p-value threshold of 0.001. In other words, we assume that a gene binds to a given transcription factor if the p-value is at most 0.001.

To assess the reliability of cluster analysis in the inference of co-regulation, we evaluated the clustering results by computing the true positive rate (TP rate), which is defined as the fraction of co-clustered gene pairs that share at least one common known transcription factor. A high TP rate indicates a high level of co-regulation from a given clustering result. As we do not have complete knowledge of all transcription factors, this TP rate is expected to be underestimated. Moreover, we compared the TP rate from a clustering algorithm to that from random partitions over a range of numbers of clusters because the TP rate may be sensitive to the number of clusters and/or the size distribution of clusters. Our primary evaluation criterion is a z-score, which measures the significance of the TP rate from a clustering result relative to the distribution of TP rates from random partitions with the same number of clusters and same cluster size distributions. Hence, the z-score is a measure of how accurately cluster analysis infers co-regulation relative to a random guess. A high z-score implies that the TP rate from the given clustering result is significantly higher than those of random partitions, and hence, indicates a high level of co-regulation.

Results

Effect of clustering algorithms

In order to study the effect of different clustering algorithms, we applied different clustering algorithms to subsets of genes listed in SCPD or YPD using all available experiments from the compendium dataset and the environmental stress dataset. For each dataset, we extracted two overlapping gene subsets according to the genes listed in SCPD and YPD respectively. For each of these four gene subsets, we evaluated the proportion of co-regulated genes from clustering results using two criteria: transcription factor databases (SCPD or YPD) and ChIP data. Because we do not have perfect knowledge of the optimal number of clusters, we applied each clustering algorithm over a range of numbers of clusters.

Two typical results are shown in Figure 2a,b, which compare the z-scores from different clustering algorithms (hierarchical average-link using correlation and Euclidean distance, hierarchical complete-link using correlation and Euclidean distance, MCLUST and IMM) on the yeast compendium dataset with 215 genes and 273 experiments evaluated using SCPD and ChIP data, respectively. The model-based clustering algorithm MCLUST with the equal-volume spherical model and hierarchical complete-link with correlation as the similarity measure produce the highest z-scores (hence, proportion of co-regulated genes) over the entire range of number of clusters (from 5 to 100), using either SCPD or ChIP data as the evaluation criterion. Figure 2a,b also shows that using correlation coefficient as the pairwise similarity measure produces significantly higher proportions of co-regulated genes from clustering results than using Euclidean distance. The results from MCLUST and IMM shown in Figure 2a,b represent z-scores from the algorithms applied to the standardized data. Standardization of the data dramatically increases the z-scores from model-based methods (see Figure A.1.b in Additional data file 1). Standardization means that the average expression value of each gene across all experiments is subtracted from the expression value of each gene and then divided by the standard deviation of its expression levels across all experiments. It can be shown that correlation and Euclidean distance are equivalent after standardization.

We observed similar results on another subset from the yeast compendium dataset, and the two gene subsets from the environmental stress data (see Figures A.2-A.4 in Additional data file 1): MCLUST with the equal-volume spherical model on the standardized data typically produces the highest z-scores and using correlation as the similarity measure always produces higher z-scores than Euclidean distance. In addition, the two independent evaluation criteria, transcription factor databases (SCPD or YPD) and ChIP data, produce very

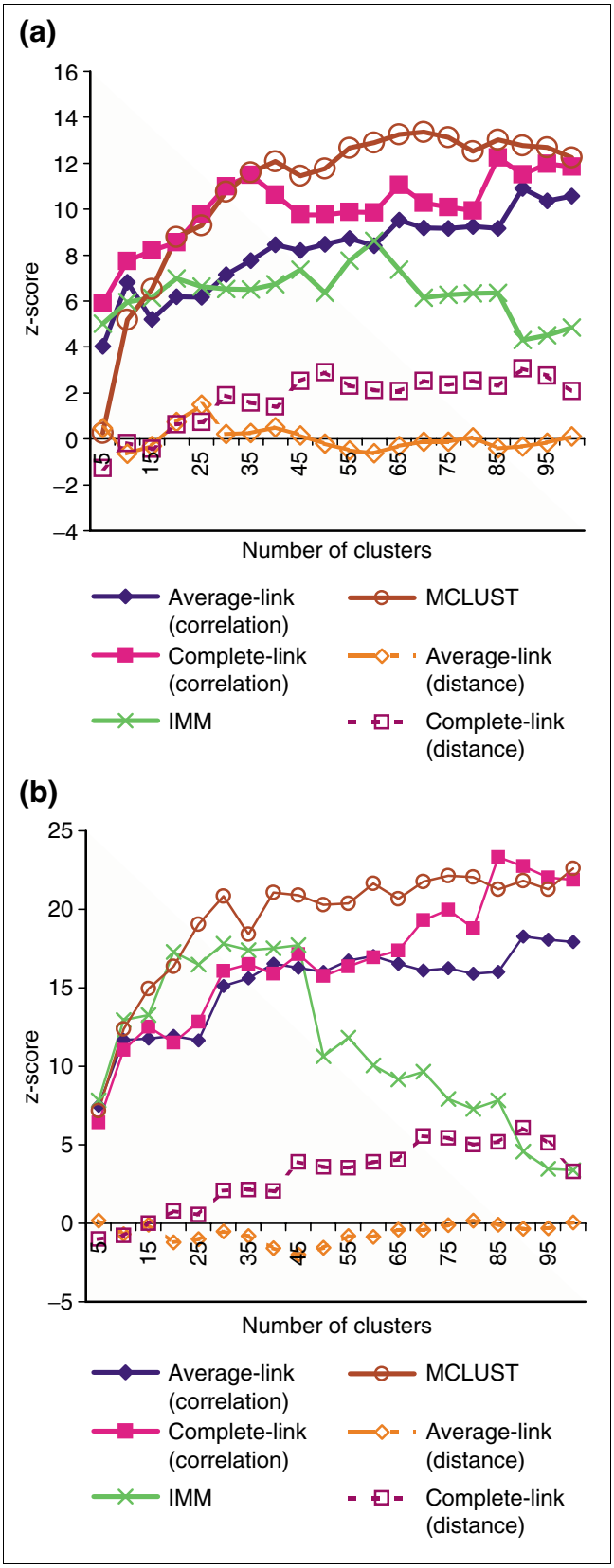

Figure 2

Figure 2
Effect of different clustering algorithms using all available microarray experiments. We compared the ability of different clustering algorithms to identify co-regulated genes using all 273 microarray experiments from the subset of compendium data with 215 genes. The clustering algorithms we compared include hierarchical average-link using correlation and Euclidean distance as the similarity measure, hierarchical complete-link using correlation and Euclidean distance as the similarity measure, and model-based clustering algorithms MCLUST and IMM on standardized data. A high z-score indicates a high proportion of co-regulated genes from clustering results compared to those from random partitions with the same numbers of clusters and cluster size distributions. Since the optimal number of clusters is not known, we compared the performance of clustering algorithms over a range of different numbers of clusters (from 5 to 100). **(a)** The transcription factor database SCPD is used as the evaluation criterion for co-regulated genes. **(b)** ChIP data is used as the evaluation criterion for co-regulated genes. The model-based clustering algorithm MCLUST produces relatively high z-scores using either SCPD or ChIP as our evaluation criterion.

similar results on the two gene subsets from both the compendium and environmental stress data.

One of the advantages of the model-based clustering methods over traditional heuristic clustering algorithms is that we can estimate the optimal number of clusters. In our previous work on model-based clustering methods [21,23], we showed that both MCLUST and IMM produced reasonable estimates of the numbers of clusters on microarray data. Using IMM, we estimated that there are 25 clusters on the compendium data subset with 215 genes and 273 experiments, 42 clusters on the compendium dataset with 537 genes and 258 experiments, 16 clusters on the environmental stress dataset with 205 genes and 205 experiments, and 34 clusters on the environmental stress dataset with 526 genes and 198 experiments. On the other hand, MCLUST does not offer any reasonable estimates of numbers of clusters in this case.

Effect of the number of microarray experiments

In order to study the effect of the number of microarray experiments on the proportion of co-regulated genes from cluster analysis on typical microarray datasets, we randomly sampled (with replacement) subsets of E experiments from each of the two gene subsets of the compendium and environmental stress data, where E = 5, 10, 20, 50, 100. We repeated this random sampling procedure 100 times for each E on each dataset. We then applied clustering algorithms to these randomly sampled subsets. Thus, we generated a clustering result for each clustering algorithm on each of these 100 randomly sampled subsets with different sizes (E). The performance of a clustering algorithm on a dataset with E experiments is summarized by the median z-score over the 100 randomly sampled subsets with E experiments.

Figure 3a shows a typical result comparing the median z-scores from a given clustering algorithm (hierarchical complete-link, in this case) on randomly sampled subsets with

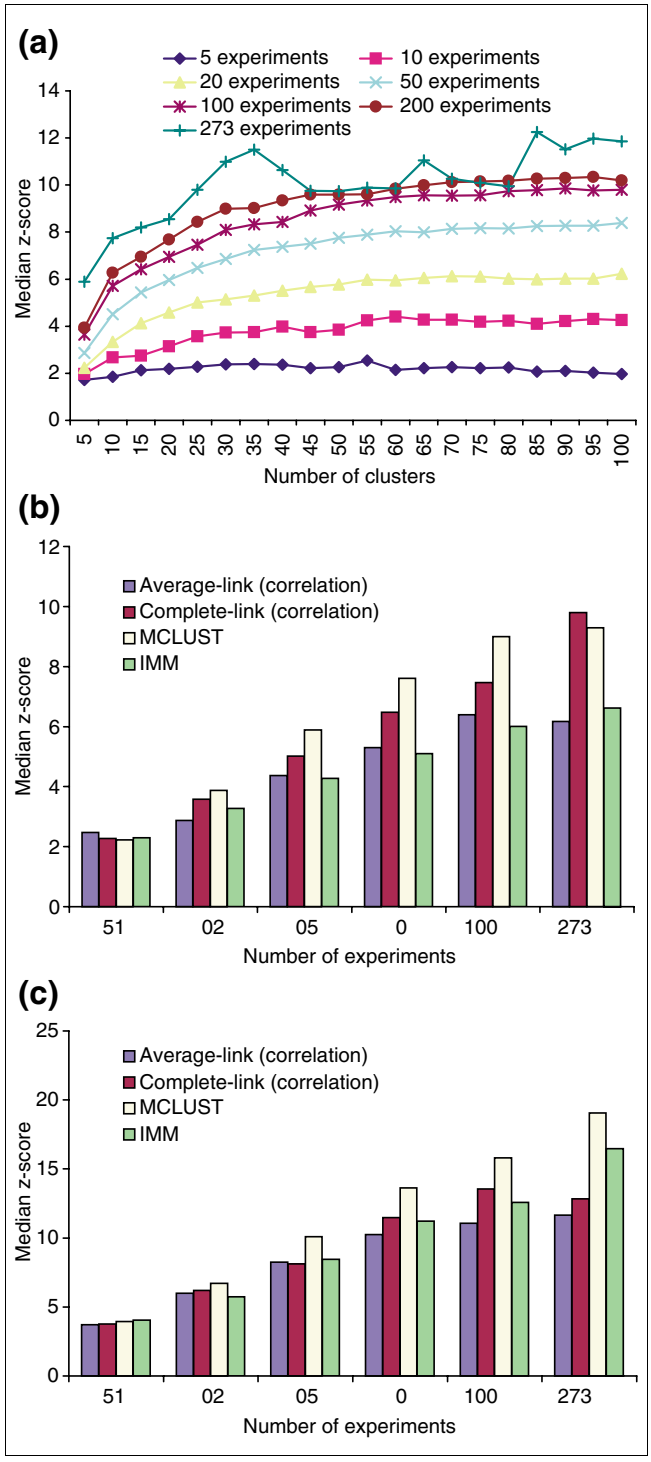

Figure 3

Figure 3
Effect of the number of microarray experiments on the compendium data subset with 215 genes. We compared the extent of co-regulated genes using different numbers of microarray experiments on the subset of compendium data with 215 genes. In order to produce typical datasets with *E* experiments (where *E* = 5, 10, 20, 50, 100), we randomly sampled (with replacement) 100 different subsets of *E* experiments from the compendium data with 215 genes and 273 experiments. The ability to identify co-regulated genes from clustering results is summarized by the median z-scores over the 100 randomly sampled datasets. A high median z-score indicates a high proportion of co-regulated genes from clustering results compared to those from random partitions. **(a)** We compared the median z-scores using different numbers of experiments (*E*) from hierarchical complete-link over a range of different numbers of clusters (from 5 to 100). The transcription factor database SCPD is used as the evaluation criterion for co-regulated genes. The median z-scores generally increase as *E* increases over different numbers of clusters. This shows that higher proportions of co-regulated genes are identified on microarray datasets with higher numbers of experiments. **(b)** Using SCPD as our evaluation criterion, we compared the median z-scores using different numbers of experiments (*E*) and different clustering algorithms (hierarchical average-link and complete-link using correlation, model-based clustering algorithms MCLUST and IMM on standardized data) on the compendium data subset with 215 genes at 25 clusters. We estimated the optimal number of clusters on this dataset to be 25 using IMM, and we observed similar results at different numbers of clusters. **(c)** Using ChIP data as our evaluation criterion, we compared the median z-scores using different numbers of experiments (*E*) and different clustering algorithms on the compendium data subset with 215 genes at 25 clusters. Using either SCPD or ChIP as our evaluation criterion, the median z-scores typically increase as *E* increases, and MCLUST typically produces relatively high median z-scores.

different numbers of microarray experiments (*E*) over a range of numbers of clusters on the compendium data subset with 215 genes evaluated using SCPD. Figure 3a shows that the median z-scores (and hence, the proportions of co-regulated gene pairs) increase as the number of microarray experiments (*E*) increases for hierarchical complete-link over different numbers of clusters. We observed the same trend using other clustering algorithms. In particular, the median z-scores increase drastically from five experiments to 50 experiments, and then the increase in median z-score starts to flatten. We observed the same trend on all our datasets (two different gene subsets from both the compendium and environmental stress data) evaluated using either a transcription factor database (SCPD or YPD) or ChIP data (see Figures B.1-B.4 in Additional data file 1 for detailed results).

Figure 3b shows a typical result comparing the median z-scores from different clustering algorithms (hierarchical average-link using correlation, hierarchical complete-link using correlation, MCLUST and IMM on standardized data) over different sizes of microarray datasets (*E* = 5, 10, 20, 50, 100, and 273) at 25 clusters on the compendium data subset with 215 genes evaluated using SCPD. Again, we observed that the median z-scores increase as the numbers of microarray experiments in the randomly sampled datasets (*E*) increase. Moreover, the model-based algorithm MCLUST produces the highest median z-scores (and hence, proportion of co-regulated gene pairs) for *E* = 5, 10, 20, 50 and 100.

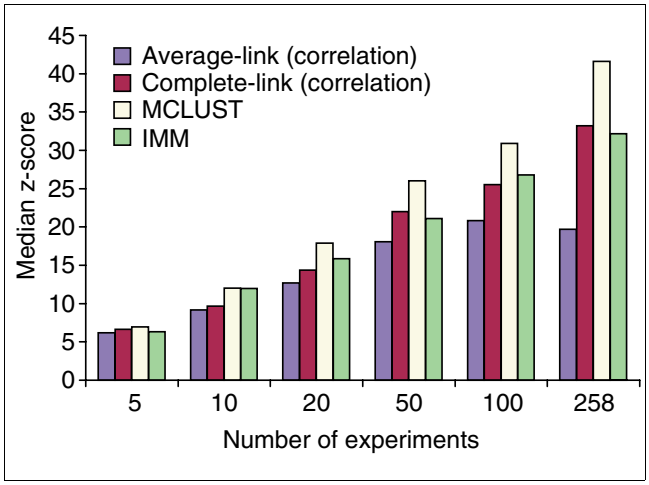

Figure 4
Effect of the number of microarray experiments on the compendium data subset with 537 genes. Using YPD as the evaluation criterion, we compared the median z-scores using different numbers of experiments and different clustering algorithms (hierarchical average-link and complete-link using correlation, model-based clustering algorithms MCLUST and IMM on standardized data) on the compendium data subset with 537 genes and 258 experiments at 40 clusters. The median z-scores (and hence, proportions of co-regulated genes) increase as the number of experiments increases, and MCLUST produces relatively high median z-scores. We observed very similar results using ChIP data as the evaluation criterion.

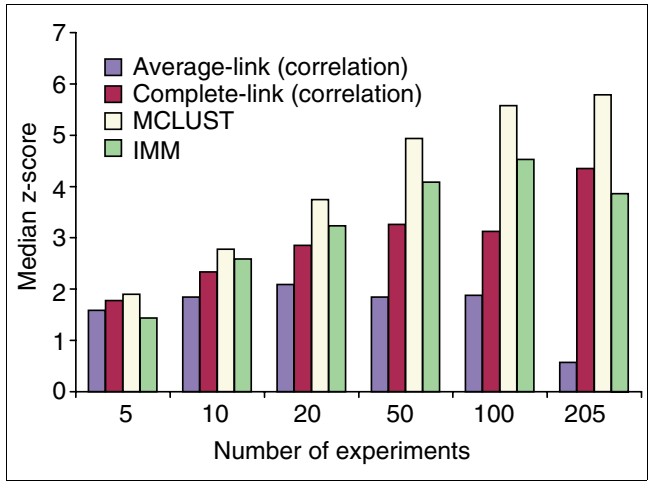

Figure 5
Effect of the number of microarray experiments on the environmental stress data subset with 205 genes. Using SCPD as the evaluation criterion, we compared the extent of co-regulated genes using different numbers of microarray experiments on the subset of environmental stress data with 205 genes and 205 experiments at 20 clusters. The median z-scores (and hence, proportions of co-regulated genes) increase as the number of experiments increases, and MCLUST typically produces relatively high median z-scores. We observed very similar results using ChIP data as the evaluation criterion.

When all the experiments are used (E = 273), hierarchical complete-link produces the highest median z-score. We observed the same trend (that is MCLUST generally produces the highest median z-scores) at other numbers of clusters.

We also compared the distribution of z-scores over the 100 randomly sampled subsets as a function of the size of the randomly sampled subsets (E). Specifically, we created box-plots of the z-scores over different sizes of randomly sampled data (E) for a given clustering algorithm and a fixed number of clusters. The medians and percentiles of the z-scores generally increase when there are more experiments in the subsets (see Figure B.1.d in Additional data file 1). In other words, a higher proportion of co-regulated genes are identified on microarray datasets with a higher number of experiments.

Using ChIP data as the evaluation criterion, Figure 3c shows a typical result comparing the median z-scores from different clustering algorithms over different E at 25 clusters on the compendium data subset with 215 genes. Both evaluation criteria (SCPD and ChIP data) produce very similar results: MCLUST generally produces the highest median z-scores and the median z-scores increase as E increases. The only difference is that MCLUST produces higher z-scores than hierarchical complete-link using correlation with all 273 experiments when ChIP data is used as our evaluation criterion instead of SCPD.

We observed the same results on all other datasets: the median z-scores increase as E increases and MCLUST generally produces the highest median z-scores compared to other clustering algorithms. For example, Figure 4 compares the performance of different clustering algorithms (hierarchical average-link using correlation, hierarchical complete-link using correlation, and MCLUST on standardized data) over different sizes of randomly sampled data (E = 5, 10, 20, 50, 100 and 258) on another gene subset from the compendium data with 537 genes evaluated using YPD. Figures 5 and 6 show the results on the two gene subsets from the environmental stress data evaluated using SCPD and YPD respectively. The z-scores of IMM are not available on all the plots because it is very computationally expensive to run IMM on 100 randomly sampled subsets for each E on each dataset. We observed the same trends when ChIP data is used as the evaluation criterion (see Figures B.1-B.4 in Additional data file 1).

Diversity of microarray experiments

We investigated the effect of the diversity of experimental conditions on the level of co-regulation from cluster analysis. Specifically, we adopted a greedy algorithm to search for a subset of E experiments with high diversity and another subset of E experiments with low diversity from each of the compendium and environmental stress datasets. For these searches, diversity was defined as average pairwise correlation in gene expression between experiments (the Materials and methods section gives details of the greedy algorithm and

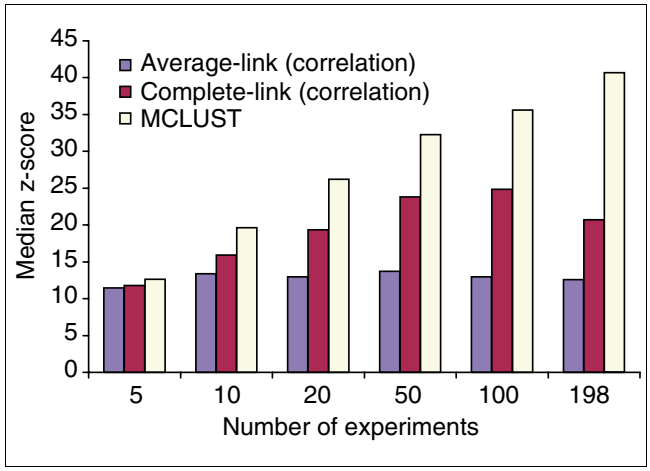

Figure 6
Effect of the number of microarray experiments on the environmental stress data subset with 526 genes. Using YPD as the evaluation criterion, we compared the extent of co-regulated genes using different numbers of microarray experiments on the subset of environmental stress data with 526 genes and 198 experiments at 30 clusters. The median z-scores increase as the number of experiments increases, and MCLUST produces relatively high median z-scores. We observed very similar results using ChIP data as the evaluation criterion.

our definition of diversity). Cluster analysis was then applied to these subsets with high and low diversities.

Contrary to our expectations, we did not observe any consistent patterns between the diversity of experimental conditions and the z-score. For example, relatively similar subsets of knock-out experiments from the compendium data tend to produce higher proportions of co-regulated genes than diverse subsets of such experiments. On the other hand, relatively diverse subsets of time-course experiments from the environmental stress data tend to produce higher proportions of co-regulated genes than similar subsets of such experiments. It is possible that the diversity of experimental conditions has a different effect on different types of microarray datasets: the compendium dataset consists of knock-out experiments, while the environmental stress dataset consists of concatenated time-course experiments. However, we need more evidence (in particular, more microarray datasets of different natures) to confirm this possibility. Another possible reason for the inconsistent patterns is that our definition of diversity in terms of average correlation between all pairs of experiments may not be the best definition of diversity for this purpose of grouping experiments that are likely to help identify co-regulated genes. A third possible reason is that the diversity of experimental conditions has no significant effect on co-regulation at all. With our current results, we cannot rule out this third possibility.

Discussion

It is important to note that even when all experiments are included, the best results produce clusters with only a 28% true positive rate (see Figure E.1.a in Additional data file 1). That is, most of the genes in a given cluster do not share a common, known transcription factor. There are several possible reasons for this. First, with the present state of knowledge, it is possible that genes in the same cluster do in fact share a common transcription factor that is not (yet) represented in the databases used as gold standards (YPD, SCPD and ChIP data). We note for example, that when one compares ChIP data to YPD, the false-negative rate is approximately 80% using the recommended *p*-value of 0.001. That is, known gene transcription factor interactions from YPD are identified only about 20% of the time by ChIP (see Table F in Additional data file 1). Hence, it is possible that our evaluation criteria all underestimate the number of co-regulated genes in a cluster. Second, gene regulation is more complex than accounted for in our approach; for example, we define sharing a common transcription factor as 'co-regulated', and each gene belongs to exactly one cluster in the clustering algorithms we considered. Individual genes are often regulated by multiple transcription factors, some of which may enhance or repress transcription. Hence, genes may be co-expressed as a result of a combination of the effects of multiple transcription factors that need not be shared across all genes. Third, genes may be included in a cluster primarily because of noise (measurement errors or technical variations) in the data rather than true signal. Finally, the range of conditions under which the experimental data was obtained may not produce changes in gene expression that would result in segregation of genes into appropriate clusters.

Even with the above caveats, our methodology simulates a common approach of experimental biologists. That is, clustering of diverse gene-expression datasets under the assumption that co-regulated genes will co-cluster, followed by attempts to identify the common transcription factors and transcription factor binding sites. While this approach has been very successful when applied to very large datasets (in particular in yeast), it is clear that the accuracy of inference should be highly dependent on the number of experimental conditions included in the analysis. The motivation of our study is to provide guidance as to the likelihood that this approach will produce true- and false-positive results and to study these rates as a function of the number of experiments that are clustered.

Our current study does not provide completely quantitative results; for example, how many experiments are sufficient for co-clustered genes to have x% probability of being co-regulated? Ideally, we would like to minimize both false positives and false negatives. However, we believe that it is of greater importance to focus on false positives, because false positives potentially lead to a waste of resources and effort to verify nonexistent relationships, while false-negatives represent

missed opportunities. We also do not provide reliable false-negative rates because of our incomplete knowledge: the transcription factor databases document known gene transcription factor interactions, but they do not give any information on known genes that are not regulated by given transcription factors. It is also not clear how to determine the *p*-value threshold for non-binding between genes and transcription factors on ChIP data. Furthermore, our study does not take the information limit of microarray data into consideration. For example, the environmental stress data consist of 225 concatenated time course microarray experiments, while the number of distinct experimental conditions is significantly less than 225. However, our results show that the ability of clustering algorithms to identify co-regulated genes increases dramatically as the number of microarray experiments used in cluster analysis is increased. In general, the ability to identify co-regulated genes in yeast datasets starts to plateau when the number of microarray experiments is greater than 50. In addition, our study indicates that the likelihood of correctly identifying co-regulated genes by clustering a small number (< 10) gene-expression experiments in yeast is quite small, and that even with large numbers of experiments, the false-positive rate may exceed the true-positive rate. For example, cluster analysis on five experiments identifies co-regulated genes only 1.5- to 6-fold more accurately than random assignment of genes to clusters on our yeast datasets (Figures 3b,c, 4 and 5). Moreover, our results were, for the most part, independent of the number of genes in the datasets and the gold standards used. That is, using SCPD, YPD, or ChIP data to identify which genes share a common transcription factor yielded very similar results in most cases. Since we extracted different (but overlapping) subsets of genes using the genes listed in SCPD and YPD for evaluation, and each of these two gene subsets are independently evaluated using ChIP data again, we have very strong confidence that our observations and general results are highly representative despite our incomplete knowledge of yeast transcription factors.

Therefore, caution is indicated before embarking on computational approaches to identify putative transcription factor binding sites in genes that co-cluster in small numbers of experiments. In addition, prudence should be exercised before embarking on expensive bench experiments to characterize these putatively identified transcription factor binding sites. Finally, it is worth noting that our current study focused on yeast, which is a simple eukaryote consisting of only 6,200 genes. We would expect that the correspondence between co-clustering and co-regulation would be lower in more complex organisms. We are interested in extending our investigation to other organisms, such as *Escherichia coli* and *Caenorhabditis elegans*, both of which have fully sequenced genomes and for which there are large microarray datasets and available resources on their transcription factors. Another possible extension to our work is the inclusion of tentative regulatory sequences as our fourth evaluation criterion. A third direction of future work would be to derive mathematically the mean and standard deviation of the distribution of the TP rates from random partitions as a function of the number of clusters and cluster sizes, so as to minimize the computational running time of our study.

Conclusions

Our results demonstrate several important overall features. First, the ability to identify co-regulated genes from co-expressed genes is strongly dependent on the number of microarray experiments used in cluster analysis, and the accuracy of these associations plateaus at between 50 and 100 experiments. Second, the model-based clustering algorithm MCLUST consistently outperforms more traditional methods in accurately assigning co-regulated genes to the same clusters. Third, using correlation as the similarity measure in heuristic-based clustering algorithms generally produces relatively higher proportions of co-regulated genes compared to Euclidean distance. Fourth, our two independent evaluation criteria for co-regulation (transcription factor databases and ChIP data) produced similar conclusions.

Materials and methods

Microarray datasets

We used two publicly available yeast microarray datasets consisting of hundreds of microarray experiments: the yeast compendium data [2,24] and the yeast environmental stress data [15,16,25,26]. The yeast compendium dataset [2] consists of 300 two-color cDNA microarray experiments in which the transcript levels of diverse mutations or chemically treated culture in yeast were compared to that of a wild-type or mock-treated cultures. The yeast environmental stress dataset [15,16] consists of 225 concatenated time course cDNA microarray experiments. These experiments represent the temporal program of gene expression in response to diverse environmental transitions (such as heat shock, hydrogen peroxide or nitrogen depletion) and to DNA-damaging agents (the methylating agent methylmethane sulfonate and ionizing radiation).

Evaluation criteria and microarray data pre-processing

We adopted two types of evaluation criteria for co-regulated genes: yeast transcription factor databases [11,12] and yeast ChIP data [13,14,27]. The SCPD [11] lists approximately 230 yeast genes that are regulated by 90 transcription factors, while the YPD [12] lists approximately 580 yeast genes that are regulated by 120 transcription factors as of November 2001. We extracted two subsets of genes from each of the compendium and environmental stress datasets: one subset of genes as listed in SCPD and another subset of genes as listed in YPD. These gene subsets are selected on the basis of the genes listed in SCPD and YPD. We did not pre-process the microarray data with any filtering steps based on differential expression or absolute levels of expression (see section G of

Table 1

Comparing YPD and SCPD

	SCPD	YPD	Common
Number of distinct ORFs	235	584	156
Number of distinct transcription factors	108	120	34
Number of gene-transcription factor interactions	473	1,056	119

Additional data file 1 for a discussion of the effect of filtering). After eliminating genes and experiments with lots of missing values, the full compendium datasets evaluated using SCPD and YPD consist of 215 genes under 273 experiments, and 537 genes under 258 experiments respectively, and the full environmental stress datasets evaluated using the SCPD and the YPD consist of 205 genes under 205 experiments, and 526 genes under 198 experiments respectively. These data subsets contain at most 1% missing expression values. As the current implementations of the model-based clustering algorithms (MCLUST and IMM) require the input data to have no missing values, we filled in the missing values using KNNimpute [28] which imputes missing values using weighted average expression levels from other genes with similar expression patterns. To our surprise, most of the gene transcription factor interactions listed in YPD are not listed in SCPD and vice versa. Table 1 shows that there are only 156 common genes listed in both YPD and SCPD, and only 119 common gene transcription factor interactions are listed in both YPD and SCPD.

Because the gene transcription factor interactions from transcription factor databases are incomplete, we independently evaluated the extent of co-regulation of these four gene subsets (two gene subsets from each of the compendium and environmental stress data) using the yeast ChIP data [14], which systematically identify target genes bound *in vivo* by a set of 106 known transcription factors. The publicly available ChIP data [14] adopts an error model in which a confidence value (*p*-value) is assigned to each regulator-DNA interaction. A *p*-value close to 0 implies that we have high confidence that a gene of interest binds to a given transcription factor. Lee *et al.* [14] recommended a *p*-value threshold of 0.001 to minimize false positives and to maximize legitimate regulator-DNA interactions.

There are 791 gene transcription factor interactions from YPD for which both the gene names and the transcription factors are present in the ChIP data [14]. Out of these 791 interactions, only 20% (or 159) were detected by the ChIP data using a *p*-value threshold of 0.001. On the other hand, there are 642 gene transcription factor interactions from the ChIP data [14] using a *p*-value threshold of 0.001. Out of these 642 interactions, only 34% (or 221) were reported in YPD. Therefore, the gene transcription factor interactions inferred from the ChIP data are quite different from those listed in YPD (see Table F in Additional data file 1).

Measure of statistical significance

We evaluated the level of co-regulation of clustering results by considering pairs of genes assigned to the same clusters and counting the fraction of these gene pairs that share at least one common known transcription factor. Specifically, we defined the true-positive rate (TP rate) as

$$\text{TP rate} = \frac{\text{Number of gene pairs from the same clusters and share at least one common transcription factor}}{\text{Number of gene pairs from the same clusters}}$$

A high true-positive rate indicates a high proportion of co-regulated genes from a given clustering result. However, as we do not have complete knowledge of all transcription factors (for example, the gene-transcription factor interactions listed in transcription factor databases are likely to be incomplete and the *p*-value threshold used in ChIP data may not be optimal), the TP rates we computed are likely to be underestimates.

As the TP rate may change as a function of the number and size distribution of clusters, we compared the TP rate from a clustering result to that from random partitions over a range of numbers of clusters. Specifically, we randomly partitioned the set of genes from a clustering result many times (typically 1,000 times in our experiments) to produce the same number of clusters and cluster size distribution as the given clustering result. We computed the TP rates of these random partitions and compared the distribution of these TP rates from random partitions to the TP rate of the given clustering result. The TP rates from random partitions typically closely follow the normal distribution. Hence, we computed the mean μ and standard deviation σ for the distribution of TP rates from random partitions. Let us denote the TP rate from a clustering result as X. The z-score, z, associated with the TP rate is defined as

$$Z = \frac{X - \mu}{\sigma}$$

A high z-score implies that the TP rate from the given clustering result is significantly higher than those of random partitions, and thus indicates a high level of co-regulation. We

computed the z-scores as a function of clustering algorithm, number of experiments and number of clusters.

Effect of the number of microarray experiments

To study the effect of the number of microarray experiments on the likelihood of discovering co-regulated genes from clustering results, we randomly sampled (with replacement) subsets of E experiments from each of the four data subsets (two gene subsets from each of the compendium and environmental stress data), where E = 5, 10, 20, 50, 100. We repeated this random sampling procedure 100 times for each E on each dataset.

Effect of clustering algorithms

There are numerous algorithms and associated programs to perform cluster analysis (for example, hierarchical methods [29], self-organizing maps [30], k-means [31], model-based approaches [17-19,32,33]) and many of these techniques have been applied to expression data (for example [1,6,20-23,33,34]). In our previous work [20,21,23], we defined cluster quality in terms of functional categories on real microarray datasets and the true underlying clusters (or classes) on synthetic datasets, and we showed that model-based clustering algorithms such as MCLUST [17-19] and the infinite mixture model-based method (IMM) [20-22] typically produced higher cluster quality than other heuristic-based clustering methods. We now focus on comparing the likelihood of assigning co-regulated genes to the same clusters from different clustering methods. In particular, we studied the performances of both heuristic-based clustering algorithms (hierarchical complete-link and hierarchical average-link) and model-based clustering algorithms (MCLUST and IMM).

Similarity measures and heuristic-based algorithms

Most heuristic-based clustering algorithms take the pairwise similarities of objects (genes or experiments) as input and create as output an organization of the objects grouped by similarity to each other. There are many similarity measures, among which the two most popular ones for gene-expression data are correlation coefficient and Euclidean distance. Correlation is a similarity measure, that is, a high correlation coefficient implies high similarity, and it captures the directions of change of two expression profiles. Euclidean distance is a dissimilarity measure, that is, a high distance implies low similarity, and it measures both the magnitudes and directions of change between two expression profiles.

Hierarchical algorithms define a dendrogram (tree) relating similar objects in the same subtrees. In agglomerative hierarchical algorithms (such as average-link and complete-link), each object is initially assigned to its own subtree (cluster). In each step, similar subtrees (clusters) are merged to form the dendrogram. We obtain clusters from the dendrogram by stopping the merging process when the desired number of clusters (subtrees) is produced. Different definitions of cluster similarity yield different clustering algorithms. In hierarchical complete-link algorithm, cluster similarity is defined to be the minimum similarity between a pair of genes, one from each of the two clusters. In hierarchical average-link algorithm, the cluster similarity of two clusters is the average pairwise similarity between genes in the two clusters.

MCLUST

The finite Gaussian mixture model-based approach assumes that each cluster follows the multivariate normal distribution, with model parameters that specify the location and property of each cluster. Different models in MCLUST assume different cluster properties (shape, volume and orientation). The most constrained model is the equal-volume spherical model in which all clusters are assumed to have equal-volume and spherical in shape. The unconstrained model is the most general model, in which the clusters can be elliptical and may have different orientations and volumes. MCLUST implements the expectation-maximization (EM) algorithm for clustering via finite Gaussian mixture models, as well as model-based hierarchical clustering algorithms, with optional cross-cluster constraints [19].

Infinite mixture model-based approach (IMM)

Medvedovic *et al.* [22] postulated an infinite Gaussian mixture model for gene-expression data which incorporates an error model for repeated measurements. Each cluster is assumed to follow a multivariate normal distribution, and the measured repeated expression levels are assumed to follow another multivariate normal distribution. They used a Gibbs sampler to estimate the posterior pairwise probabilities of co-expression. These posterior pairwise probabilities are treated as pairwise similarities, which are used as inputs to clustering algorithms like hierarchical complete-link algorithm. Our recent work showed that applying hierarchical complete-link to these posterior pairwise probabilities using a particular cluster similarity parameter (minimum distance = 0.9999) yields reasonable estimates of the optimal number of clusters [21]. In this work, IMM produced very reasonable estimates of the optimal numbers of clusters for all our datasets.

Diversity of microarray experiments

We defined the diversity of a set of E microarray experiments as the average correlation of all pairs of experiments in this set of E experiments. A high correlation implies low diversity, while a low correlation implies high diversity. To investigate the effect of diversity of experimental conditions on the level of co-regulation, we used a greedy algorithm to select a set of E experiments with high diversity (low average correlation) and another set of E experiments with low diversity (high average correlation) from the compendium and environmental stress data subsets, where E = 5, 10, 15, 20, 30, 40, 50, 80, 100.

Let S be the current set of experiments. In the case of searching for a set of E highly diverse experiments, we initialized S with the pair of experiments with minimum pairwise

correlation. After the initialization step, the following two steps are repeated until there are E experiments in S. First, search for an experiment e with minimum total correlation to all the current experiments in S; then, add experiment e to the set S. The greedy algorithm to search for a set of E experiments with low diversity is very similar, except that experiments with maximum correlation are added instead.

Additional data files

A pdf file (Additional data file 1) available with the online version of this article gives the datasets used in this work. The files and software are also available from our website [35].

Acknowledgements

We thank Professor Elton T Young from Biochemistry at University of Washington, members of the Bumgarner lab and the working group organized by Adrian Raftery at University of Washington for their feedback and comments for this project. K.Y.Y. and R.E.B. are supported by NIH-NIDDK grant 5U24DK058813-03. R.E.B. is also supported by NIH-NIAID grants 5P01 AI052106-02, 1R21AI052028-01 and 1U54AI057141-01, NIH-NIEHA grant 1U19ES011387-02, NIH-NIDA grant 1 P30 DA015625-01, NIH-NHLBI grants 5R01HL072370-02 and 1P50HL073996-01. M.M. is supported by NHGRI grant 1R21HG002849-01.

References

1. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Nat Acad Sci USA* 1998, **95:**14863-14868.
2. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, *et al.*: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102:**109-126.
3. Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R, Altschuler SJ: **Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters.** *Nat Genet* 2002, **31:**255-265.
4. Wyrick JJ, Young RA: **Deciphering gene expression regulatory networks.** *Curr Opin Genet Dev* 2002, **12:**130-136.
5. 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.
6. Tavazoie S, Huges JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22:**281-285.
7. Geiss GK, Carter VS, He Y, Kwieciszewski BK, Holzman T, Korth MJ, Lazaro CA, Fausto N, Bumgarner RE, Katze MG: **Gene expression profiling of the cellular transcriptional network regulated by alpha/beta interferon and its partial attenuation by the hepatitis C virus nonstructural 5A protein.** *J Virol* 2003, **77:**6367-6375.
8. Ohler U, Niemann H: **Identification and analysis of eukaryotic promoters: recent computational approaches.** *Trends Genet* 2001, **17:**56-60.
9. Wolfsberg TG, Gabrielian AE, Campbell MJ, Cho RJ, Spouge JL, Landsman D: **Candidate regulatory sequence elements for cell cycle-dependent transcription in *Saccharomyces cerevisiae*.** *Genome Res* 1999, **9:**775-792.
10. Jelinsky SA, Estep P, Church GM, Samson LD: **Regulatory networks revealed by transcriptional profiling of damaged *Saccharomyces cerevisiae* cells: Rpn4 links base excision repair with proteasomes.** *Mol Cell Biol* 2000, **20:**8157-8167.
11. Zhu J, Zhang MQ: **SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*.** *Bioinformatics* 1999, **15:**607-611.
12. Costanzo MC, Hogan JD, Cusick ME, Davis BP, Fancher AM, Hodges PE, Kondu P, Lengieza C, Lew-Smith JE, Lingner C, *et al.*: **The yeast proteome database (YPD) and *Caenorhabditis elegans* proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information.** *Nucleic Acids Res* 2000, **28:**73-76.
13. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, *et al.*: **Genome-wide location and function of DNA binding proteins.** *Science* 2000, **290:**2306-2309.
14. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, *et al.*: **Transcriptional regulatory networks in *Saccharomyces cerevisiae*.** *Science* 2002, **298:**799-804.
15. 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.
16. 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.
17. Fraley C, Raftery AE: **MCLUST: Software for model-based cluster analysis.** *J Classification* 1999, **16:**297-306.
18. Fraley C, Raftery AE: **How many clusters? Which clustering method? - Answers via model-based cluster analysis.** *Computer J* 1998, **41:**578-588.
19. Fraley C, Raftery AE: **Model-based clustering, discriminant analysis, and density estimation.** *J Am Stat Assoc* 2002, **97:**611-631.
20. Yeung KY, Medvedovic M, Bumgarner RE: **Clustering gene expression data with repeated measurements.** *Genome Biol* 2003, **4:**R34.
21. Medvedovic M, Yeung KY, Bumgarner R: **Bayesian mixture model based clustering of replicated microarray data.** *Bioinformatics* 2004, **20:**1222-1232.
22. Medvedovic M, Sivaganesan S: **Bayesian infinite mixture model based clustering of gene expression profiles.** *Bioinformatics* 2002, **18:**1194-1206.
23. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL: **Model-based clustering and data transformations for gene expression data.** *Bioinformatics* 2001, **17:**977-987.
24. **Functional discovery via a compendium of expression profiles** [http://www.rii.com/publications/2000/cell_hughes.html]
25. **Genomic response of yeast to diverse stress conditions** [http://genome-www.stanford.edu/yeast_stress]
26. **Genomic responses to DNA-damaging agents** [http://www-genome.stanford.edu/Mec1]
27. 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.
28. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: **Missing value estimation methods for DNA microarrays.** *Bioinformatics* 2001, **17:**520-525.
29. Hartigan JA: *Clustering Algorithms* New York: Wiley; 1975.
30. Kohonen T: *Self-organizing maps* Berlin: Springer-Verlag; 1997.
31. MacQueen J: **Some methods for classification and analysis of multivariate observations.** In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* Edited by: Cam LML, Neyman J. Berkeley: University of California Press; 1965:281-297.
32. McLachlan GJ, Basford KE: *Mixture Models: Inference and Applications to Clustering* New York: Marcel Dekker; 1988.
33. McLachlan GJ, Bean RW, Peel D: **A mixture model-based approach to the clustering of microarray expression data.** *Bioinformatics* 2002, **18:**413-422.
34. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Nat Acad Sci USA* 1999, **96:**2907-2912.
35. **UW Department of Microbiology - bioinformatics publications** [http://expression.washington.edu/publications/kayee/coregulation]