Discriminative Clustering of Yeast Stress Response

Samuel Kaski[1,2], Janne Nikkilä[1], Eerika Savia[1] and Christophe Roos[3]

[1] Neural Networks Research Centre, Helsinki University of Technology
P.O. Box 5400, FIN-02015 HUT, Finland
{samuel.kaski, janne.nikkila, eerika.savia}@hut.fi
[2] Department of Computer Science, University of Helsinki, Finland
[3] Medicel Ltd., Helsinki, Finland
christophe.roos@helsinki.fi

Summary. When a yeast cell is challenged by a rapid change in the conditions, be it temperature, osmolarity, pH, nutrient or other, it starts a genome stress response program. Survival of especially single-cell organisms depends on their ability to adapt to the environmental changes and therefore stress response has received much attention. In the budding yeast *Saccharomyces cerevisiae* several hundred genes out of about 6500 present in the genome have previously been found involved in a stereotyped stress response pattern. Hierarchical clustering techniques applied to gene expression measurements have also previously identified a subset of genes termed common environmental stress response (CESR) or common environmental response (CER) genes, that respond in the same way in a variety of environmental conditions. There is evidence from two different sets of experiments that many of these genes are regulated by the same Msn2p and Msn4p transcription factor pair. We have extended the study by in silico data mining using a new supervised discriminative clustering (DC) technique, which directly searches for responses potentially regulated by the Msn2/4p factors. We observed a cluster of CESR/CER genes, comparable to those previously found and potentially regulated by Msn2/4p. The results of discriminative clustering both support the viability of the technique in supervised gene expression clustering and yield new insights into genomic stress response.

1 Introduction

The ability of the yeast *Saccharomyces cerevisiae* genome to respond to environmental changes is vital, since no cellular condition of gene activity is universally optimal. The response of yeast cells to stress induced by drastic changes in the environment has been used as a paradigm to study gene regulation networks. It is also important to understand the cell response to stress to its own value, since virtually any treatment introduces some kind of stress situation for the yeast cells, and is thus present in any gene activity measurement. Moreover, understanding yeast gene regulation will help as a model

for studies on higher organisms. While it is clear that understanding gene regulation requires data on chromosome structure, gene activity (transcriptome), protein pool to mention only the major concepts, the transcriptome has received the most attention due to the high throughput measurement technologies available (gene chip/microarray).

The gene expression of the yeast under stress has been studied extensively [1, 3, 9, 13], and it has become evident that a certain group of the yeast genes is always activated during various stress treatments. The genes in this set are often called *common environmental response (CER)* genes [1], or *environmental stress response, ESR* genes [3]. In this paper we adopt the term from [1], and call them CER genes.

Due to differences in the documentation of the experiments in [3] and [1], it still is somewhat uncertain whether the group of CER genes found in one experimental setting is the same as the set of genes found in the other experiments. Even more unclear is the understanding of the regulatory system of the yeast stress response. There seem to be at least a few general "stress regulators" like Yap1p, Msn2p, and Msn4p, that are shown to be required for a large set of CER genes to be induced [1, 3]. In addition, the existence of condition specific regulators, like Hsf1p for heat shock, has been noted [9].

We carry out a meta-analysis to study the concurrence of the two different CER gene definitions and the two independent sets of measurements. Additionally, we refine *in silico* the earlier analyses of the role of the Msn2p and Msn4p transcription factors in regulating the CER genes. We use a new statistical data mining tool called *discriminative clustering (DC)* [7, 15, 16] which differs from standard clustering by being supervised by class labels of the data.

The clusters in DC partition the data into mutually similar sets, in the same way as in the standard K-means clustering. The difference is that DC maximizes the dependence of the clusters and the classes. An intuitive description of what DC does is that it uses the classes as hints on which samples should be considered similar. Samples should be more similar if they belong to the same class; more precisely, distances in directions where the class distribution changes more should be larger.

In this study the classes are chosen according to the response of the strain lacking *Msn2/4p* to a stress treatment. Then DC will consider genes more similar if their response is the same even after the potential stress regulator Msn2/4p is removed. The cluster analysis becomes more focused on regulation by Msn2/4p, instead of taking all differences in gene activation into account.

2 Discriminative Clustering

Consider a set of paired data $(\mathbf{x}, c)$, where $\mathbf{x} \in \mathbb{R}^n$ are continuous-valued multivariate observations of primary data and c are discrete classes. In this work each $\mathbf{x}$ is a profile of expression of a yeast gene in various stress treatments.

In a nutshell, we wish to find clusters of $\mathbf{x}$ that are maximally dependent on c. This task has two parts. (i) In order to call the data groups *clusters*, they need to be local in the primary data space, that is, contain similar expression profiles. The second part is that (ii) the clustering should capture the dependency between the primary data and the classes.

The motivation for (i) is that even though the clusters are supervised, they can still be interpreted in the same way as "normal clusters" in unsupervised clustering, as sets of similar data. The motivation for (ii) is that choosing the classes properly allows us to focus the analysis to the variation relevant to the classes. In this work we want to find evidence for regulation by Msn2/4p, and we choose the classes to show how the genes react to stress treatments after the Msn2/4p has been removed. Maximization of dependency with the classes then forces the clustering to focus on similarities in the expression profiles that are relevant to regulation by Msn2/4p. Genes regulated in the same way will become more similar.

2.1 Definition of Clusters

Each cluster j is defined by a prototype $\mathbf{m}_j$. Samples $\mathbf{x}$ are assigned to the clusters that have the closest prototype: $\mathbf{x}$ belongs to cluster j if $\|\mathbf{x} - \mathbf{m}_j\| \leq \|\mathbf{x} - \mathbf{m}_k\|$ for all k. Here the distance is the standard Euclidean distance. This definition is the same as in the standard K-means clustering method, for instance.

2.2 Measuring Dependency

The clusters and the classes form a contingency table, a cross tabulation of the two categorizations of the same data. The count of data n_{ji} within cell (j, i) tells how many samples of class i occur in the cluster j. The margin $n_{j\cdot} = \sum_i n_{ji}$ gives the number of samples within cluster j, and the fixed margin $n_{\cdot i} = \sum_j n_{ji}$ gives the total number of samples in class i.

The dependency between the clusters and the classes can be measured based on the contingency table. If the true proportion of data occurring within each cell, i.e. the joint distribution p_{ji}, was known, the dependency could be measured by mutual information. However, since only a finite sample is available, the mutual information computed from the empirical distribution would be a biased estimate. A Bayesian finite-data alternative is the *Bayes factor* between models that assume dependent and independent margins. Bayes factors have classically been used as dependency measures for contingency tables (see, e.g., [4]). We have used the classical results as building blocks to derive the Bayes factor to be optimized; the novelty in DC is that we suggest maximizing the Bayes factor instead of only measuring dependency of fixed tables with it.

2.3 The Cost Function

In general, frequencies over the cells of a contingency table, as well as over the margins, are multinomially distributed. The model M_i of *independent margins* assumes that the multinomial parameters of the contingency table cells are determined by the posterior parameters at the margins. In the alternative model M_d of *dependent margins*, the cell-wise frequencies are assumed to have been sampled directly from a multinomial distribution over the whole contingency table, which indirectly determines the margins. Dirichlet priors are assumed for both the margin and the table-wide multinomials.

Maximization of the Bayes factor

$$\mathrm{BF} = \frac{p(\{n_{ji}\}|M_d)}{p(\{n_{ji}\}|M_i)} \tag{1}$$

with respect to the clusters then gives a contingency table where the margins are maximally dependent, that is, which cannot be explained as a product of independent margins. The cluster margin is determined by the distribution of the learning data set into the clusters, and the clusters in turn are defined by their parameters (the cluster prototypes $\mathbf{m}_j$). The BF is maximized with respect to the parameters.

After marginalization over the multinomial parameters, the Bayes factor, assuming a fixed class margin, takes the form [16]

$$\mathrm{BF} = \frac{\prod_{ji} \Gamma(n_{ji} + n^0)}{\prod_j \Gamma(n_{j\cdot} + N^0)} \; . \tag{2}$$

Here $n_{j\cdot} = \sum_i n_{ji}$ is the cluster margin, that is, number of data samples in the clusters, and the parameters n^0 and $N^0 = \sum_i n^0$ come from the Dirichlet priors. We have set $n^0 = 1$.

For large data sets compared to the number of clusters, (2) is approximated by mutual information of the margins. Another interesting connection, shown in [16], is that the Bayes factor equals the posterior density $p(\{\mathbf{m}\}|D)$ of the set of the cluster parameters $\{\mathbf{m}_j\}$ of a certain predictive model. The model predicts the class distribution within each cluster with a multinomial distribution.

2.4 Optimization

The difficulty in optimizing (2) is that the data counts n_{ji} within the clusters are discontinuous functions of the values of the cluster parameters $\mathbf{m}_j$. The counts change only when a data point changes from one cluster to another. Hence, the derivatives of the cost function are always either zero or undefined.

We have used a heuristic smoothing technique to make gradient-based optimization possible. It has worked about as well as the theoretically better justified simulated annealing that is much heavier computationally [7]. The

"number" of samples is smoothed by $n_{ji} = \sum_{c(\mathbf{x})=i} y_j(\mathbf{x})$, where $c(\mathbf{x})$ is the class of $\mathbf{x}$ and $y_j(\mathbf{x})$ is a smoothed cluster "membership function", defined by $y_j(\mathbf{x}) = Z(\mathbf{x})^{-1} \exp(-\|\mathbf{x} - \mathbf{m}_j\|^2/\sigma^2)$ with Z such that $\sum_j y_j(\mathbf{x}) = 1$, and σ governing the degree of smoothing. The standard conjugate gradient algorithm was used for the optimization. The smoothing is used only during optimization; afterwards the clusters partition the data space.

2.5 Related Methods

Discriminative clustering is closely related to the Information Bottleneck principle [2, 17] and distributional clustering [12]. The main difference is that in DC the primary data $\mathbf{x}$ is continuous-valued whereas in distributional clustering it has always been categorical. For continuous-valued data the clusters need to be defined and parameterized as partitions of the data space, which makes the algorithms and solutions very different. Although no algorithm has been developed it could in principle be possible to use the Information Bottleneck definition for continuous data as well. Then the clusters would not be local, however, and hence not as easily interpretable as "normal clusters".

Another line of related work is model-based clustering of the joint distribution of the data [5, 11]. The difference is that DC as such does not model the margin $p(\mathbf{x})$ at all; it is a predictive model of the conditional density $p(c|\mathbf{x})$. The motivation for this choice comes from the learning metrics principle [6, 8] which uses the classes c to derive a Riemannian distance measure to the primary data space. In the new metric the class distribution changes homogenously, which stretches the directions where the class distribution changes rapidly and contracts the directions where it does not. This is the desired result if changes in the class distribution are the interesting thing in the data. The metric can and has been used to supervise a variety of standard data analysis methods. The connection to DC is that it can be shown [6] that under restrictive assumptions DC is asymptotically equivalent to standard K-means in such a metric.

A direct connection between modeling of joint density and DC is that by including a model for $p(\mathbf{x})$, DC can be regularized to a model of joint density, $p(\mathbf{x}, c) = p(\mathbf{x})p(c|\mathbf{x})$. If $p(c|\mathbf{x})$ comes from standard DC and $p(\mathbf{x})$ from a K-means type model, then the cost function of DC becomes a tunable compromise between K-means and DC [7]. This compromise can be interpreted as regularization of DC towards K-means, which is useful for small data sets, assuming the density structure in $p(\mathbf{x})$ contains useful hints for the prediction task.

3 Data

Both Causton and colleagues [1] and Gasch and colleagues [3] have used DNA micro-arrays to analyze changes in the transcriptome (pool of all gene tran-

scripts from a cell or cell population) in yeast cells responding to a panel of diverse environmental stresses. The conditions include treatment with heat, changes in pH, in salt concentration, in osmolarity, in reactive oxygen or nutrient concentrations. In each condition, first a reference time-point is measured and then transcriptome data from a set of consecutive time points following the environmental challenge are gathered. Altogether, each of the two research groups has gathered about 150 micro-array measurements covering the full yeast genome of about 6200 known or predicted genes. Both groups attempt to define "genomic expression programs", in other words groups of genes that are commonly involved in handling most or all stress challenges.

In [1] a set of "Common Environmental Response" (CER) genes is defined as follows: First genes, whose expression was found to be induced or repressed in all conditions, were identified by visual inspection from a hierarchically organized tree. Then the genes that changed at least twofold (up or down) in five or six time courses were selected as CER genes. The authors collected 499 genes with a common response to most of the environmental changes examined. Of the 499 genes, 216 were found up-regulated (activated) while 283 were down-regulated (repressed). In [3] the environmental stress response (ESR) genes are more strictly defined: two hierarchical clusters of genes, one with ca 300 activated, the other with ca 600 repressed genes were identified as having a stereotyped response to each of the stress conditions. In all the time series the data was divided by the value of the respective timepoint zero.

It is known from previous studies that many stress response genes are under the regulation of the Msn2p and Msn4p transcription factors [10] and therefore both groups also make attempts to measure stress response in yeast strains mutant for these transcription factors. In [1] a CER subset (not documented) of 136 genes is identified based on their opposite behavior in the mutant lacking *Msn2/4p* as compared to the wild type control in the acid challenge experiment. Since we were not able to obtain this list of 136 genes, we imitated the original preprocessing of the data according to the documentation, resulting in a set of 4146 genes, which we analyze further with DC.

In [3] an ESR subset of 180 genes depending on Msn2/4p or the Yap1p transcription factors is documented. In order to compare these findings with ours we had to find matching genes in the data set of [1]. A match was found for a subset of 143. We will refer to this common set of genes by "dependent CER genes from [3]".

4 Results

4.1 Msn2/4p Regulated CER Genes by Discriminative Clustering

In [1], the CER genes were identified and analyzed in two stages. First, it was assumed that the CER genes react in the same way in all of the stressful environmental conditions. The expression profiles of all genes were clustered,

and sets of the most up- and down-regulated genes were identified by visual inspection from the hierarchical clustering tree. Second, the response of the genes to a mutation in the putative regulators, Msn2/4p, was studied. A set of genes up-regulated in the wild type but down-regulated in the mutant strains was identified as CER genes potentially regulated by Msn2/4p.

This setting is perfectly suited for discriminative clustering. The goal is to cluster the expression profiles to discover similarly behaving genes. Yet, pure unsupervised clustering is not enough; it is particularly interesting to search for those similarities in expression that are regulated by Msn2/4p.

In the discriminative clustering setup, the expression profile of a gene is $\mathbf{x}$, and the supervising class label c comes from the response of the gene to the mutation. We quantized the response to acid treatment after mutation (vs. time-point zero) to three classes: **down**: strongly down-regulated after mutation (one quarter of genes); **up**: strongly up-regulated (one quarter of genes); and **no change**: the rest. DC then finds clusters of genes that (1) behave similarly in the set of stress treatments and (2) respond similarly to the mutation.

We start by verifying the technical findings quantitatively, by checking that the dependencies the supervised clustering finds are replicable. Then we interpret the results and compare the findings qualitatively with those in [1]. Due to differences in reporting of the results in the papers, quantitative comparison is possible only with [3]; it will be carried out in Sect. 4.2.

DC Results Are Replicable

In order to verify that the results of supervising the clustering are real and not merely results of overfitting the clusters to noise in the data, we compared them to standard unsupervised K-means clusters with cross-validation.

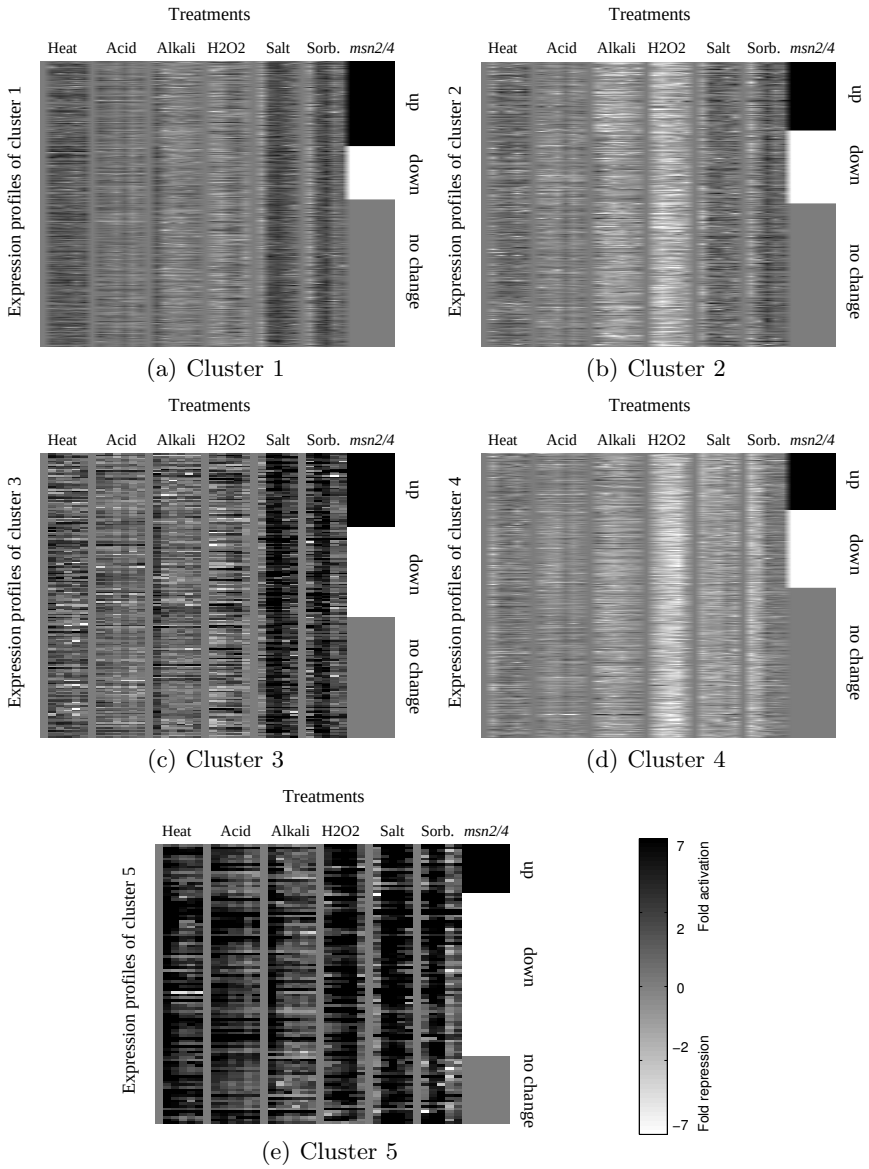
The smoothing parameter $\sigma = 0.9$ used in the optimization was chosen with a validation set in preliminary experiments, and the number of clusters was set (heuristically) to 12. DC was initialized by K-means.

In the cross-validation study the data was randomly divided into $N = 20$ sets. Clusters were computed with $N-1$ of the sets, and the results evaluated with the remaining test set. T-test over the $N = 20$ replications showed that the DC consistently ($P < 0.001$) found dependencies between the classes and the expression profiles. The performance measure was (2).

DC Finds a CER Cluster Downregulated in Mutant Lacking Msn2/4p

The expression profiles of the yeast genes in the 12 clusters are shown in **Figs. 1.–3.** DC was computed of the whole set of 4146 genes used in [1]. Each expression profile is a set of time series under different stress treatments. The time labels are shown in **Fig. 4.** and a detailed description of the time series can be found in [1].

The most striking finding is the cluster number 5 (enlarged in **Fig. 4.**), containing 103 genes that have highly upregulated expression in all the treatments for the wild strain and an exceptionally large proportion of the genes are



(a) Cluster 1 (b) Cluster 2

(c) Cluster 3 (d) Cluster 4

(e) Cluster 5

Fig. 1. Gene expression profiles in a set of stress treatments clustered into 12 DC clusters. Each subfigure contains the genes within one cluster, and each row is the profile of one of the genes. The rightmost column shows the class of the gene: whether it is up- or down-regulated in the acid treatment in the mutant strain lacking *Msn2/4p*. Continued in **Figs. 2.–3.**

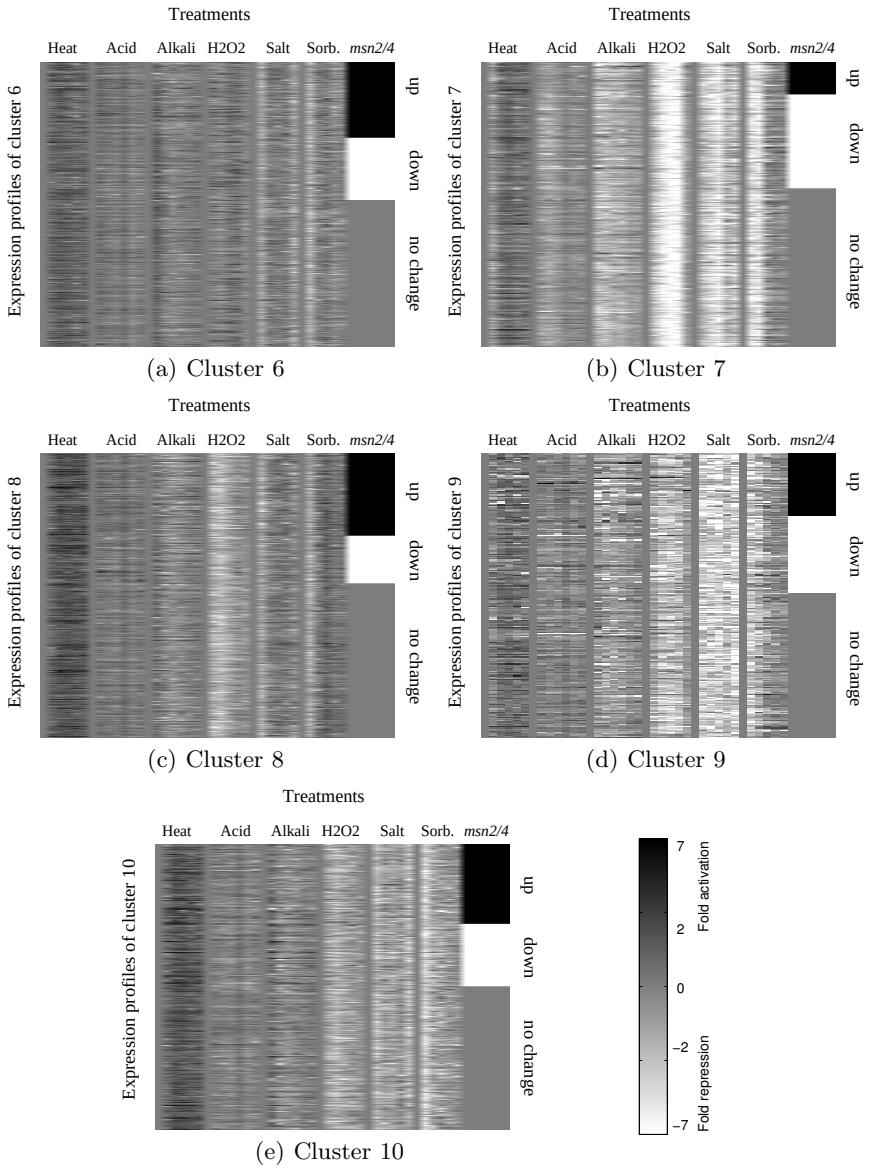


Fig. 2. Gene expression profiles in a set of stress treatments clustered into 12 DC clusters. Each subfigure contains the genes within one cluster, and each row is the profile of one of the genes. The rightmost column shows the class of the gene: whether it is up- or down-regulated in the acid treatment in the mutant strain lacking *Msn2/4p*. Continued in **Figs. 1.** and **3.**

downregulated in the mutant strain lacking *Msn2/4p* (**Fig. 5.**). This cluster is the most likely candidate for CER genes that are regulated by Msn2/4p.

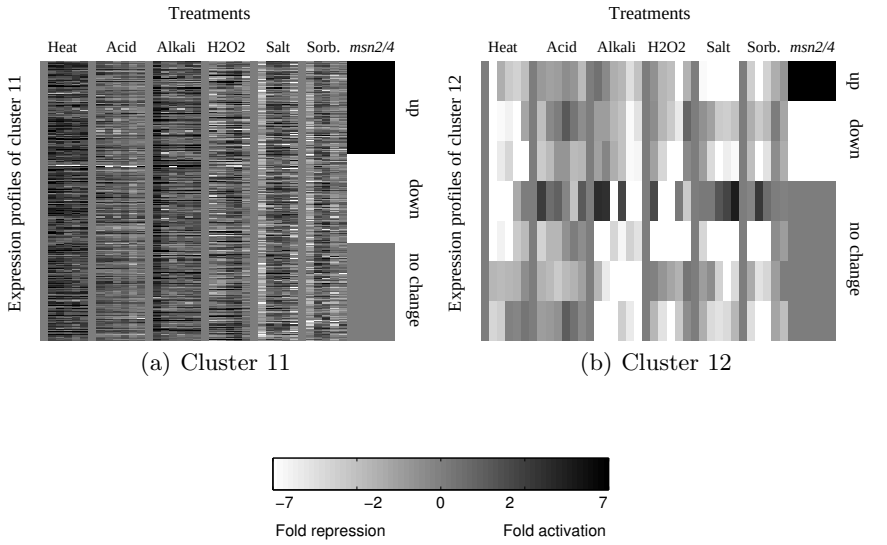
The possibility that the class distribution within cluster number 5 could have arisen by chance was evaluated by random sampling. If there is no interaction between the classes and the clusters, the distribution of data in the contingency table is determined completely by the distribution of data in its margins, that is, the classes and clusters. We sampled a large set (10,000) of contingency tables under the hypothesis that the margins are independent, and estimated for each contingency table cell how unexpected the observed value is. The P-value for obtaining a more extreme value than the observed number of samples was computed as a percentage within the sampled set.

The resulting P-values for each contingency table cell are shown in **Tables 1.** and **2.**. For cluster number 5 the number of downregulated genes is much larger than expected ($P < 0.001$) and the number of non-affected genes is much lower than expected ($P < 0.001$). Hence, it is very unlikely that the observed interaction of the effect of *msn2/4* mutation and the very active CER-type response profile of the genes would have arisen by chance.

Table 1. Unexpectedness of the enriched contingency table cells. The table shows P-values for those cells where the number of samples exceeded the expected amount. For instance, in cluster 5 the number of downregulated genes is significantly higher than expected, whereas the number of upregulated and not changed genes is smaller than expected (marked by "–" and treated in **Table 2.**)

	upregulated	downregulated	no change
Cluster 1	0.02	–	0.31
Cluster 2	–	0.41	0.46
Cluster 3	0.40	0.07	–
Cluster 4	–	0.17	0.23
Cluster 5	–	**< 0.01**	–
Cluster 6	0.22	–	0.31
Cluster 7	–	**< 0.01**	0.11
Cluster 8	0.06	–	0.13
Cluster 9	–	0.30	0.43
Cluster 10	0.15	–	0.47
Cluster 11	0.01	0.03	–
Cluster 12	–	0.44	0.45

We cannot verify quantitatively how closely our findings match those of Causton *et al.* [1] since they do not report the full list of gene names. We will, however, compare our list with the list of another study [3] in Sect. 4.2.



(a) Cluster 11 (b) Cluster 12

Fig. 3. Gene expression profiles in a set of stress treatments clustered into 12 DC clusters. Each subfigure contains the genes within one cluster, and each row is the profile of one of the genes. The rightmost column shows the class of the gene: whether it is up- or down-regulated in the acid treatment in the mutant strain lacking *Msn2/4p*. Continued from **Figs. 1.–2.**

Table 2. Unexpectedness of the contingency table cells with diminished number of samples. The table shows P-values for those cells where the number of samples falls below the expected amount. For instance, in cluster 5 the number of not changed genes is significantly smaller than expected, whereas the number of downregulated genes is larger than expected (marked by "–" and treated in **Table 1.**)

	upregulated	downregulated	no change
Cluster 1	–	$<$ **0.01**	–
Cluster 2	0.38	–	–
Cluster 3	–	–	0.11
Cluster 4	0.02	–	–
Cluster 5	0.08	–	$<$ **0.01**
Cluster 6	–	0.08	–
Cluster 7	$<$ **0.01**	–	–
Cluster 8	–	$<$ **0.01**	–
Cluster 9	0.20	–	–
Cluster 10	–	0.13	–
Cluster 11	–	–	$<$ **0.01**
Cluster 12	0.31	–	–

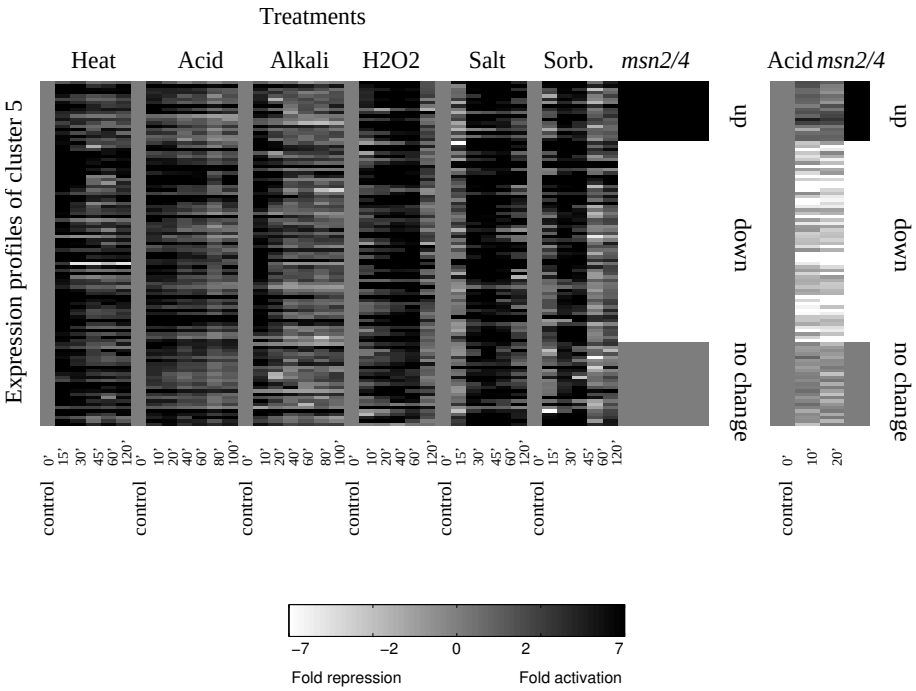
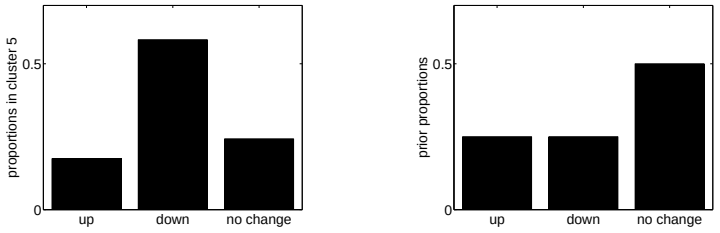


Fig. 4. Left: Enlarged gene expression profiles in a set of stress treatments for genes in cluster 5. Each row is the profile of one of the genes. Right: Expression profiles of the same genes in the mutant strain lacking *msn2/4*. These profiles have been used for defining the classes of the genes (shown in the rightmost columns). The classes tell whether the genes are up- or down-regulated in the acid treatment in the mutant.



(a) Distribution of the genes of cluster number 5 in the classes

(b) Distribution of genes of the whole genome

Fig. 5. Behavior of the genes of cluster number 5 in the mutant strain lacking *msn2/4* differs markedly from the expected behavior

Other Findings

We tried to see if the clustering would find a group of stress response genes, the expression of which is independent of the Msn2/4p regulation. Cluster 7 (**Fig 1 (b)**) contains many down-regulated genes, especially in the peroxide and osmotic shock experiments. Since Msn2/4p are primarily transcription activator factors (and not repressors) [10] these down-regulated genes are probably not under the direct control of Msn2/4p. However, these genes could be under an indirect control of Msn2/4p if one considers that Msn2/4p could activate some secondary repressing regulators. Therefore, the relatively high abundance of down-regulated genes in cluster 7 is not in itself a reflection of Msn2/4p-independence. Now, if the down-regulation of these genes would be indirectly repressed by Msn2/4p, the expression should rise in the mutant strain lacking Msn2/4p. Interestingly, this does not seem to be the case for a fairly ($p<0.01$, **Table 1**) large amount of these genes that remain down-regulated also in the mutant. Therefore, we conclude that cluster 7 might contain a significant amount of stress response genes independent of the Msn2/4p-regulation.

However, not all clusters with stress-activated genes obey the Msn2/4p regulation, as can be seen in cluster number 1. **Tables 1.** and **2.** reveal additional potentially interesting interactions between the expression profiles and the Msn2/4p regulation. The genes in cluster number 1 are predominantly up-regulated within the stress treatments (see **Fig. 1.**) but only very few of them are affected by the mutation. Hence, they are likely regulated by some other transcription factors than Msn2p or Msn4p.

4.2 DC Findings Are Consistent with Experiments in a Different Stress Treatment

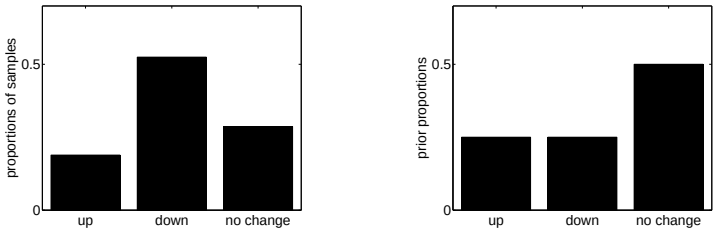
Two groups [1, 3] have sought for yeast stress-induced genes and their regulation by Msn2/4p. The main difference is that the former studied the response of the *msn2/4* mutant in acid stress and the latter in hydrogen peroxide and heat stress. The independent sets of measurements were made with different measurement techniques (cDNA microarrays vs. Affymetrix chips). If the genes are true CER genes they should of course react generally to any type of stress, and hence be equally detectable in either set of experiments.

So far in this article we have only used the measurements of one of the groups [1]. Now the results of the other group will be used in an independent evaluation to verify our findings. Since replication studies are relatively scarce in large-scale gene expression studies because of the cost of the measurements, it will additionally be interesting to see how consistent the findings from the two data sets are. Our study provides some indirect evidence on this.

The Findings Are Consistent

As a sanity check, we first compared whether the set of CER-type genes found to be down-regulated in the *Msn2/4p* mutants in the independent study [3]

were down-regulated in [1] as well. A matching gene was found for a subset of 143 genes; we will refer to this set as "dependent CER genes from [3]". Within this set, exceptionally many genes are down-regulated in the independent measurements of the acid treatment [1] as well (**Fig. 6.**). The distribution differs significantly (chi-square test, $P < 0.001$) from the expected distribution estimated from the whole data set, which completes our sanity check.



(a) Distribution of dependent CER genes from [3]

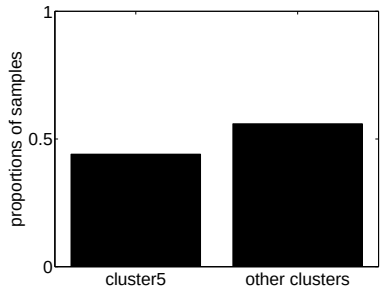
(b) Expected distribution

Fig. 6. Behavior of dependent CER genes from [3] in the acid treatment of [1] (**a**) differs strongly from the overall expected behavior computed from all the genes in [1] (**b**)

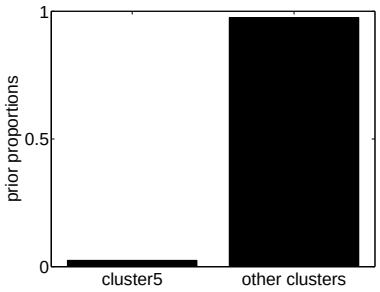
Next, we used the results of the independent study to verify the DC results. Based on Figure 5, the cluster number 5 should contain a large proportion of CER genes regulated by Msn2/4p. This finding is based on analyzing one of the data sets with DC, and now the result is compared with a non-DC analysis of the other independent set. If the result is favourable, it will support the viability of DC.

Fig. 7. shows that the proportion of the independently found CER genes in cluster number 5 is exceptionally high; the number differs significantly from chance (chi-square test, $P < 0.001$).

Not all of the dependent CER genes found in [3] belong to cluster 5, however. Nevertheless, they are distributed very inhomogenously in the DC clusters (**Fig. 8.**). In particular, a number of them have ended up in clusters 1, 3, and 11. These clusters contained the largest proportion of generally up-regulated genes in the DC-clusters (**Figs 1.–3.**). In clusters 1 and 11 the behavior of the mutant strains differed clearly from chance (**Tables 1.** and **2.**). This suggests that some genes predicted to be Msn2/4p-regulated end up in the different clusters because they are not solely dependent on Msn2/4p. Indeed, regulation of gene transcription in yeast, as in other organisms, is achieved by synergistic binding between several transcription factors and other proteins building up the transcription initiation complex.

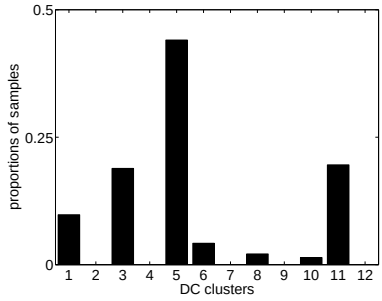


(a) Proportion of dependent CER genes from [3] in cluster 5 vs. other clusters

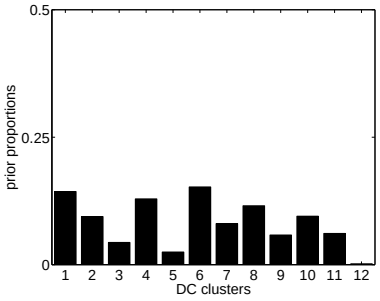


(b) Expected proportions of data in cluster 5 vs. other clusters

Fig. 7. The dependent CER genes found in an independent study [3] are considerably enriched in cluster 5 (**a**), compared to the expected number of genes calculated from all the genes in [1](**b**)



(a) Distribution of dependent CER genes from [3] on DC clusters



(b) Prior distribution of data on DC clusters

Fig. 8. The dependent CER genes found in an independent study [3] are concentrated on only a few DC clusters (**a**), compared to the distribution expected based on the whole data from [1](**b**)

5 Discussion

In summary, we have applied a new supervised clustering method, discriminative clustering (DC), to mine gene expression profiles for common environmental response (CER) genes and their regulatory mechanisms.

The clustering was supervised to focus on gene expression relevant to regulation by certain transcription factors, Msn2/4p. The findings are consistent with both of the two earlier studies on the same problem [1, 3]. DC has been originally developed for supervised mining of large data sets, and the results support its usefulness in genome-wide mining of expression data.

Additionally, the DC clustering suggested possible subclasses within the set of CER genes.

The DC complements standard unsupervised clustering by making it possible to supervise the exploration of data. Ultimately, when the resulting hypotheses mature, they need to be tested with even more focused methods and models. The current findings suggest a follow-up study where the stress-induced genes that are down-regulated in mutants lacking *Msn2/4p* mutants would be sought directly by searching for genes with high activity (up-regulation) in the wild type and low activity (down-regulation) in the mutants.

As a side study, we compared indirectly the results of two research groups, working with different methods and published in different papers [1, 3]. To the extent the documentation allows, the results seemed compatible. It would be interesting to continue the present DC study by generalizing from one supervisory signal, class labels, to multiple classifications derived from the response of the mutant strains to different stress treatments. Data is already available by [3]. The results should reveal more about the compatibility of the different data sets and should yield more accurate hypotheses about which genes are true CER-genes and respond similarly to all kinds of stress treatments. Moreover, instead of quantizing the responses to three classes they could be considered as multivariate continuous-valued observations. Then the recent generalization of discriminative clustering from categorical supervisory signal to continuous-valued multivariate signal [14] could be the proper data analysis tool.

Acknowledgments

This work was supported by the Academy of Finland, grants 50061 and 52123.

References

1. Helen C. Causton, Bing Ren, Sang Seok Koh, Christopher T. Harbison, Alanita Kanin, Ezra G. Jennings, Tong Ihn Lee, Heather L. True, Eric S. Lander, and Richard A. Young. Remodeling of yeast genome expression in response to environmental changes. *Molecular Biology of Cell*, 12:323–337, 2001.
2. Nir Friedman, Ori Mosenzon, Noam Slonim, and Naftali Tishby. Multivariate information bottleneck. In *Proc. Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 152–161. Morgan Kaufmann Publishers, San Francisco, CA, 2001.
3. Audrey P. Gasch, Paul T. Spellman, Camilla M. Kao, Orna Carmel-Harel, Michael B. Eisen, Gisela Storz, David Botstein, and Patrick O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11:4241–4257, 2000.
4. I. J. Good. On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Annals of Statistics*, 4:1159–1189, 1976.

5. Trevor Hastie and Robert Tibshirani. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society B*, 58:155–176, 1996.
6. Samuel Kaski and Janne Sinkkonen. Principle of learning metrics for data analysis. *The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology, Special Issue on Data Mining and Biomedical Applications of Neural Networks*, accepted for publication.
7. Samuel Kaski, Janne Sinkkonen, and Arto Klami. Regularized discriminative clustering. In Christophe Molina, Tülay Adali, Jan Larsen, Marc Van Hulle, Scott Douglas, and Jean Rouat, editors, *Neural Networks for Signal Processing XIII*, pages 289–298. IEEE, New York, NY, 2003.
8. Samuel Kaski, Janne Sinkkonen, and Jaakko Peltonen. Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 12:936–947, 2001.
9. W. H. Mager and A.J. De Kruijiff. Stress-induced transcriptional activation. *Microbiological Reviews*, 59:506–531, 1995.
10. M. T. Martinez-Pastor, G. Marchler, C. Schuller, A. Marchler-Bauer, H. Ruis, and F. Estruch. The saccharomyces cerevisiae zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE). *EMBO Journal*, 15:2227–2235, 1996.
11. David J. Miller and Hasan S. Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. In Michael Mozer, Michael Jordan, and Thomas Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 571–577. MIT Press, Cambridge, MA, 1997.
12. Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of English words. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 183–190. ACL, Columbus, OH, 1993.
13. H. Ruis and C. Schuller. Stress signaling in yeast. *Bioessays*, 17:959–965, 1995.
14. J. Sinkkonen, J. Nikkilä, L. Lahti, and S. Kaski. Associative clustering by maximizing a bayes factor. Technical Report A68, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 2003.
15. Janne Sinkkonen and Samuel Kaski. Clustering based on conditional distributions in an auxiliary space. *Neural Computation*, 14:217–239, 2002.
16. Janne Sinkkonen, Samuel Kaski, and Janne Nikkilä. Discriminative clustering: Optimal contingency tables by learning metrics. In T. Elomaa, H. Mannila, and H. Toivonen, editors, *Proceedings of ECML'02, 13th European Conference on Machine Learning*, pages 418–430, Berlin, 2002. Springer.
17. Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *37th Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377, Urbana, Illinois, 1999.