International Journal of Neural Systems, Vol. 15, No. 4 (2005) 237–246

World Scientific
www.worldscientific.com

EXPLORATORY MODELING OF YEAST STRESS RESPONSE AND ITS REGULATION WITH gCCA AND ASSOCIATIVE CLUSTERING

JANNE NIKKILÄ
Laboratory of Computer and Information Science
Helsinki University of Technology
P.O. Box 5400, FI-02015 HUT, Finland
janne.nikkila@hut.fi

CHRISTOPHE ROOS
Medicel Oy, Huopalahdentie 24, FI-00350 Helsinki, Finland
christophe.roos@helsinki.fi

EERIKA SAVIA
Laboratory of Computer and Information Science
Helsinki University of Technology
P.O. Box 5400, FI-02015 HUT, Finland
eerika.savia@hut.fi

SAMUEL KASKI*
Department of Computer Science, University of Helsinki
P.O. Box 68, FI-00014 University of Helsinki, Finland
and
Laboratory of Computer and Information Science
Helsinki University of Technology
P.O. Box 5400, FI-02015 HUT, Finland
samuel.kaski@hut.fi

We model dependencies between m multivariate continuous-valued information sources by a combination of (i) a generalized canonical correlations analysis (gCCA) to reduce dimensionality while preserving dependencies in $m-1$ of them, and (ii) summarizing dependencies with the remaining one by *associative clustering*. This new combination of methods avoids multiway associative clustering which would require a multiway contingency table and hence suffer from curse of dimensionality of the table. The method is applied to summarizing properties of yeast stress by searching for dependencies (commonalities) between expression of genes of baker's yeast *Saccharomyces cerevisiae* in various stressful treatments, and summarizing stress regulation by finally adding data about transcription factor binding sites.

Keywords: Associative clustering; canonical correlation analysis (CCA); exploratory data analysis; gene expression; yeast stress.

1. Introduction

Integration of multiple information sources is becoming increasingly important in bioinformatics. It has become evident that most of the important questions in molecular and cell biology cannot be answered by studying a single data source, like DNA sequence or gene expression. Due to complexity and noise of biological systems, either lots of prior knowledge or data are required to constrain models, and single sources of measurements then suffice only for producing

*Corresponding author.

general overviews. In contrast, approaches that combine several relevant data sources have shown potential to answer specific biological questions.[4,22] However, the works integrating heterogeneous data sets presented so far tend to be tailored to specific tasks. Their application to other problems is usually possible only after tedious tailoring by both methodological and application area experts. Typical examples of these methods are Bayes nets[7] and inductive logic programming.[17]

We investigate machine learning methods that can both integrate multiple information sources and at the same time are generally applicable to a set of problems of a certain kind. We assume that the multiple data sets are multivariate and continuous-valued. The problem is to find what is common in the data sets, commonality being defined as such properties of the data sets that are consistently (statistically) dependent. The question is then how to utilize this kind of multiple information in a principled way to search for dependencies with minimal assumptions about their nature. Specifically, we study yeast stress response on gene expression level, and its regulation by a set of proteins called *transcription factors*.[21]

Yeast stress response has been studied intensively during recent years.[5,9] A group of genes appears to be always affected in various stress treatments, and this set has been called *common environmental response* (*CER*) or *environmental stress response* (*ESR*) genes. Such stress response is practically the only way for the simple yeast to respond to various adverse environmental conditions, and because it is so easy to elicit, it has been used as a paradigm to study gene regulatory networks. Even this response is far from being completely understood, however; different studies do not even agree on the sets of stress genes.

In practice we have available a set of gene expression profiles, that is, a time series of expression for each gene, for each stress treatment. The common environmental response should be visible as some properties that are unknown but common to all treatments. We assume here that the dependencies between our gene expression data sets are rather simple, perhaps even linear, since the data sets are relatively low-dimensional and all the treatments should induce rather similar expression patterns in yeast. Based on this we assume that the interesting information between expression data sets can be found by searching for global common variation between the data sets.

We will maximize the variation that is common to the stress treatment data sets and try to neglect all the other variation in the gene expression profiles. This problem can be solved by a form of *generalized canonical correlation analysis*.[1] Its interpretation as a mutual information-maximizing method further justifies the use of that specific variant of generalized canonical correlation analysis.

To explore regulation of the stress response, we further search for commonalities with data about how likely different transcription factors (TF), regulators of gene expression, are to bind to the promoter regions of the genes. If a set of genes has commonalities in their expression patterns across stress treatments, and furthermore commonalities in the binding patters, they are likely to be stress response genes regulated in the same way.

This kind of dependency exploration can be done by maximizing the mutual information, or preferably its finite-data variant, between the data sets. We will use a previously introduced method, called *associative clustering* (*AC*),[14] which clusters two data sets by maximizing dependency between the clusterings, and hence should suit the task perfectly. In practice, a linear method scales better to multiple data sets, and hence we use generalized canonical correlations as a preprocessing method to reduce the number of data sets.

Clustering methods that maximize mutual information between the data sets have been formalized in the information bottleneck framework,[8,25] directly applicable to discrete data, and extended to continuous vectorial data.[3,13] Associative clustering can be viewed as an extension to the current information bottleneck algorithms for finite sets of continuous-valued data.

The methods of clustering with constraints are also close in spirit to AC; there the clustering is supervised by constraining pairs of samples to belong (or not to belong) to the same cluster.[2,26] This will lead to similar results as IB-type clustering if samples from the same class are constrained to belong to the same cluster. Common to AC is that this can be interpreted as one data set (the constraints) supervising the other. In AC two data sets supervise each other, in the sense that the goal is to find dependencies between them.

Graphical models of dependencies between variables are another related popular formalism. They have been applied to modeling regulation of gene expression.[7] The main practical difference from our clustering approach, which makes the models complementary, is that our clusterings are intended to be used as general-purpose, data-driven but not data-specific, exploratory tools. Associative clustering is a multivariate data analysis tool that can be used in the same way as standard clusterings to explore regularities in data. The findings can then be dressed as prior beliefs or first guesses in the more structured graphical models, hopefully helping to restrict the hopelessly large search space of possible model structures.

The technical difference from standard clusterings, as well as from standard graphical models, is that our objective function is maximization of dependency between the data sets. Instead of modeling all variability in the data the models focus on those aspects that are common in the different data sets, in the sense of being consistently dependent. This fits perfectly the present application.

2. Methods

2.1. *Dependencies by mutual information*

Having observed for one set of objects (genes) several multivariate variables $V_1, \ldots, V_M, X$ (stress treatments and TF-binding) forming several data sets, our aim is to find clusters of genes that maximize the mutual information $\mathrm{I}(V_1; \ldots; V_M; X)$, or its finite-data version, that is, the Bayes factor. In principle, AC could be extended to search for a multiway contingency table between the multiple data sets, but this would lead to severe estimation problems with a finite data set.

Noting that

$$\mathrm{I}(V_1; \ldots; V_M; X) = \mathrm{I}(V_1; \ldots; V_M) + \mathrm{I}((V_1, \ldots, V_M); X)$$

we propose a sequential approximation: first approximate $\mathrm{I}(V_1; \ldots; V_M)$ by forming the optimal representation $Y(V_1, \ldots, V_M)$ with gCCA, then maximize $\mathrm{I}(Y; X)$ with AC. In this way we can reduce our problem to the AC of two variables and, in a sense, preserve dependencies between all the data sets in a computationally tractable way. Additionally, note that we are here specifically interested in *clusters* of genes, which justifies the use of AC instead of using only gCCA which merely produces a projection of the data.

2.2. *Generalized canonical correlation analysis*

We focus on the variation that is common to two or more of the M data sets and are willing to lose information that is due to variables within one data set only. For this purpose Canonical Correlation Analysis (CCA) is a natural choice. While Principal Component Analysis (PCA) works with a single random vector and maximizes the variance of projections of the data, CCA works with a pair of random vectors and maximizes the correlation between sets of projections. While PCA leads to an eigenvalue problem, CCA leads to a generalized eigenvalue problem.

There are several ways to generalize canonical correlation analysis to more than two sets of variables.[1,16] Here we use a generalization of CCA (gCCA) that has a simple connection to mutual information.[1] The gCCA problem can be written as a generalized eigenproblem

$$C\xi = \lambda D\xi \tag{1}$$

where C is the covariance matrix of the concatenated data and D is a block diagonal matrix that consists of the within-set covariance matrices of the individual data sets C_i.

In effect, CCA whitens the covariance matrices within each data set and then performs PCA to seek the largest variance in the between-data-set covariances.

2.3. *Information-theoretic interpretation*

The gCCA projection can also be interpreted from an information-theoretic point of view. Assuming that the variables are normally distributed, there is a simple connection between mutual information and CCA[18]:

$$\mathrm{I}(V_1; V_2) = -\frac{1}{2} \ln\left(\frac{\det C}{\det C_1 \det C_2}\right) = -\frac{1}{2} \prod_i \lambda_i = -\frac{1}{2} \prod_i (1 - \rho_i^2), \tag{2}$$

where the λ_i are the eigenvalues of Eq. (1) and further, where the ρ_i are the canonical correlations. For multiple data sets the equation generalizes to

$$\begin{aligned}\mathrm{I}(V_1;\ldots;V_M) &= \sum_{i=1}^{M} H(V_i) - H(V_1,\ldots,V_M)\\ &= -\frac{1}{2}\ln\frac{\det C}{\det C_1\cdots\det C_M}. \end{aligned} \quad (3)$$

We now show how the whitening of the original data sets preserves the mutual information. Using the notation of Eq. (1), such a whitening can be written as

$$V' = D^{-1/2}V. \quad (4)$$

The covariance matrix C' of the transformed variable V' is

$$C' = \mathbb{E}[D^{-1/2}VV^TD^{-1/2}] = D^{-1/2}CD^{-1/2}, \quad (5)$$

so the mutual information is preserved in the transformation:

$$\begin{aligned}\mathrm{I}(V_1';\ldots;V_M') &= -\frac{1}{2}\ln\det C'\\ &= -\frac{1}{2}\ln\frac{\det C}{\det C_1\cdots\det C_M}\\ &= \mathrm{I}(V_1;\ldots;V_M). \end{aligned} \quad (6)$$

Moreover, after data-set-wise whitening the mutual information depends only on the joint entropy, which can be shown as follows. In general, mutual information can be written as

$$I(V_1;\ldots;V_M) = \sum_{i=1}^{M} H(V_i) - H(V_1,\ldots,V_M). \quad (7)$$

The entropy of an individual data set V_i with dimensionality d_i is in general

$$\begin{aligned}H(V_i) &= -\int p(v_i)\ln p(v_i)\,dv\\ &= \frac{d_i}{2}\left(\ln(2\pi)+1\right) + \frac{1}{2}\ln\det C_i. \end{aligned} \quad (8)$$

After whitening of the within-data covariances the covariance-dependent term of the entropy vanishes,† giving $H(V_i') = \frac{d_i}{2}(\ln(2\pi)+1)$.

The mutual information $I(V_1';\ldots;V_M')$ is now

$$I\big(V_1';\ldots;V_M'\big) = \text{const.} - H\big(V_1',\ldots,V_M'\big), \quad (9)$$

which intuitively means that all the mutual information is now represented by the joint entropy plus a constant.

Our goal is to make a dimensionality reduction that maximally preserves the mutual information $I\big(V_1';\ldots;V_M'\big)$. Thus, according to Eq. (9) the best approximation is the one that maximally preserves the entropy, $H\big(V_1',\ldots,V_M'\big)$. The dimensionality is reduced by sequentially searching for the one-dimensional projection that best approximates the original d-dimensional variable. We thus seek for the one-dimensional projection $V'' = a^TV'$, with $a^Ta = 1$, that maximizes the entropy of the projection, $H(V'')$. The entropy of the projected variable V'' is

$$H(V'') = \frac{1}{2}\ln\mathrm{var}\ V'' + \text{const.}, \quad (10)$$

where

$$\mathrm{var}\ V'' = \mathbb{E}[a^TV'(V')^Ta] = a^TC'a \quad (11)$$

and therefore, the entropy of the projection will be

$$H(V'') = \frac{1}{2}\ln a^TC'a + \text{const.} \quad (12)$$

Since $a^TC'a$ is the variance of V' in the direction of a, it is maximized by choosing a to be the first principal component of C'. So, actually, the maximization of the entropy coincides with maximization of the variance, i.e., PCA for the transformed variable V'. This is equivalent to performing gCCA for the original variable V, since, as already noted in the previous section, gCCA produces the same result as whitening the within-data covariances of the data sets and then performing PCA.

2.4. *Associative clustering*

Having observed a set of paired data samples $\{(\mathbf{x}_k,\mathbf{y}_k)\}_k$ from two continuous, vector-valued random variables X and Y, we wish to find subsets of data that are informative of dependencies between X and Y. We use a previously introduced method, associative clustering (AC),[14] that produces two sets of partitions, one for X and the other for Y. The aim of AC is to make the cross-partition contingency table represent as much of the dependencies between

†One should note that whitening (i.e., transforming the covariance matrix to the identity matrix) is not the only possibility to set $\frac{1}{2}\ln\det C_i = 0$, but it is the only one that implies equal contributions of the original variables within each data set. Because we have no prior information that any of the variables within a data set would be more important than the others, it seems reasonable to require equal variances.

X and Y as possible. A Bayesian criterion for the dependency is detailed below.

Because the partitions for $\mathbf{x}$ and $\mathbf{y}$ define the margins of the contingency table, they are called *margin partitions*, and they split the data into *margin clusters*. The cells of the contingency table correspond to pairs of margin clusters, and they split data pairs $(\mathbf{x}, \mathbf{y})$ into clusters. Denote the count of data within the contingency table cell on row i and column j by n_{ij}, and sums with dots: $n_{i\cdot} = \sum_j n_{ij}$.

The AC optimizes the margin partitions to maximize dependency between them, measured by the *Bayes factor* between two hypotheses: the margins are independent (H) vs. dependent ($\bar{H}$). The Bayes factor is (derivation omitted)

$$\frac{P(\{n_{ij}\}|\bar{H})}{P(\{n_{ij}\}|H)} \propto \frac{\prod_{ij} \Gamma(n_{ij} + \alpha_{ij})}{\prod_i \Gamma(n_{\cdot i} + \alpha_i) \prod_j \Gamma(n_{j\cdot} + \alpha_j)}, \tag{13}$$

where the α come from priors, set equal to 1 in this work. The cell frequencies are computed from the training samples $\{(\mathbf{x}_k, \mathbf{y}_k)\}_k$ by mapping them to the closest margin cluster in each space. The clusters are parameterized by prototype vectors $\mathbf{m}$.

During optimization the partitions are smoothed to facilitate the application of gradient-based methods. As the final cost function to optimize, we have

$$\begin{aligned} \log BF' = &\sum_{ij} \log \Gamma\left(\sum_k g_i^{(x)}(\mathbf{x}_k) g_j^{(y)}(\mathbf{y}_k) + \alpha_{ij} \right) \\ &- \lambda^{(x)} \sum_i \log \Gamma\left(\sum_k g_i^{(x)}(\mathbf{x}_k) + \alpha_i \right) \\ &- \lambda^{(y)} \sum_j \log \Gamma\left(\sum_k g_j^{(y)}(\mathbf{y}_k) + \alpha_j \right), \end{aligned} \tag{14}$$

where

$$g_i^{(x)}(\mathbf{x}) \equiv Z^{(x)}(\mathbf{x})^{-1} \exp\bigl(-\|\mathbf{x} - \mathbf{m}_i^{(x)}\|^2 / \sigma_{(x)}^2\bigr),$$

and similarly for $g^{(y)}$. The $g(\cdot)$ are the smoothed Voronoi regions at the margins. The $Z(\cdot)$ is set to normalize $\sum_i g_i^{(x)}(\mathbf{x}) = \sum_j g_j^{(y)}(\mathbf{y}) = 1$. The parameters σ control the degree of smoothing of the Voronoi regions and the λ are regularizing parameters: if set larger than one, they favour solutions with equal-sized margin clusters. AC is optimized with conjugate gradient method. More details can be found from the previous publications.[14]

2.5. *Uncertainty in results*

Our use of Bayes factors in AC is different from their traditional use in hypothesis testing.[10] We do not test any hypotheses but the Bayes factor is maximized to explicitly hunt for dependencies. However, for the current implementation of AC, this leaves the Bayes factor of AC conditioned on the clustering model. Together with the finiteness of the real world data and local minima in optimization, this results in uncertainty in the results.

We tackle the problem of uncertainty by using bootstrap[6] to produce several perturbed clusterings. There are analogous approaches in the literature, where a more traditional clustering method has been applied to gene expression data, and has been bootstrapped.[15] In our case, we wish to find *cross-clusters* (contingency table cells) that signify dependencies between the data sets, and that are reproducible.

Reproducibility of the clusters is estimated with bootstrap, by sampling 100 bootstrap data sets from the original data set and clustering each with AC. Dependency of clusters in each bootstrap AC is estimated by generating several (1000) data sets of the same size as the original one from the marginals of the contingency table (i.e., under the null hypothesis of independence). Cross-clusters containing more observations than expected by chance given the independent margins ($p < 0.01$ with Bonferroni correction) are defined as dependent.

The two criteria of dependency and reproducibility will finally be combined by evaluating, for every gene pair, how likely they are to occur within the same significantly dependent cross-cluster in clustering models computed in the different bootstrap data sets. This similarity matrix will finally be summarized by hierarchical clustering. Cutting the tree (arbitrarily) then gives the the most dependent, robust subsets of the data.

2.5.1. *Contributions of the original variables in clusters*

Finally, the obtained reliable clusters are to be interpreted in terms of the original variables. In other words, we investigate which transcription factors bind strongly in a specific cluster. Additionally, the binding tendency should of course be reliable.

We utilize here the localness of our clusters in each data space and compute the average profile of the original data for the final TF-binding clusters. Average profiles are compared to randomized average profiles, computed from the data vectors for 10000 random sets of the same size as the cluster. This offers a way to identify abnormally high and small values of TF-binding in the cluster.

3. Yeast Gene Expression under Stress, and Its Regulators

3.1. *Data*

We used data from several experiments to analyze the dependencies between yeast stress genes and their regulators. Common stress response was sought from expression data of altogether 16 stress treatments[5,9]: heat (2), acid, alkali, peroxide, NaCl, sorbitol(2), H2O2, menadione, dtt(2), diamide, hypo-osmotic, aminoacid starvation, and nitrogen depletion. A short time series had been measured from each, and in total we had 104 dimensions. For these genes we picked up the TF-binding profiles of 113 transcription factors,[20] to search for dependencies with expression patterns. In total we ended up having 5998 yeast genes. All the values were normalized with respect to the zero point of each time series (or other control), and then the natural logarithm of these ratios was taken. Missing values were imputed by genewise averages in each data set.

3.2. *Dimensionality Reduction by gCCA*

We started with the 16 separate stress genes expression data sets, in total 104-dimensional expression data. The number of gCCA components was chosen such that the same components could be found from left-out data reasonably well (measured with the angle between the components) in 20-fold crossvalidation. This resulted in 12 generalized canonical components.

We then checked whether gCCA managed to produce meaningful components. This was verified by testing the association of the 12 first gCCA components to genes known to be affected by stress, namely the putative environmental stress response genes (ESR) found earlier.[9] Of the 12 generalized canonical components 9 showed statistically significant association to ESR genes known to be either up-regulated or down-regulated (Wilcoxon rank-sum test; $p < 0.05$). gCCA thus managed to capture the variation relevant to ESR genes.

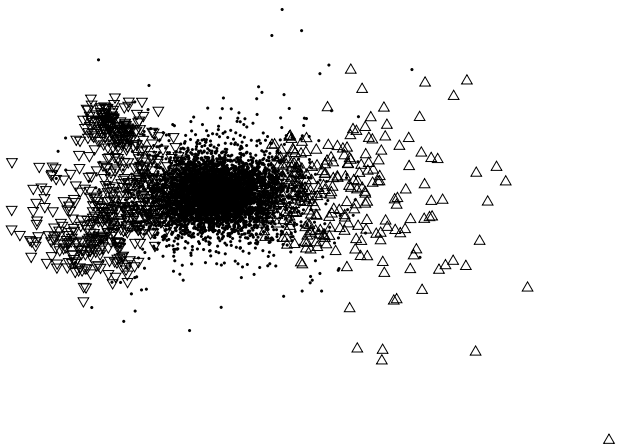

Fig. 1. Projection of the genes on the first two gCCA components revealing how the known ESR genes are separated from the rest of the genes by gCCA. Triangle up: upregulated ESR genes, triangle down: downregulated ESR genes, dots: the rest of the genes.

Figure 1 demonstrates that the putative ESR genes are separated reasonably well from the gene mass even in a two-dimensional projection.

3.3. *Associative clustering for stress expression and TF binding*

We next analyzed with associative clustering the dependencies between the 12-dimensional expression data, resulting from preprocessing by gCCA, and the 113-dimensional TF-binding data. The goal was to find subsets of genes having maximal dependency between their TF-binding and expression under stress. The number of AC clusters was chosen to produce about 10 data points in the cross-partition table (equivalently, contingency table) on the average, resulting in a table with 30×20 cells.

To verify that there are dependencies the AC is able to detect, the contingency table produced by AC was compared with the contingency table produced by independent K-means clusterings in both the data spaces, in a 10-fold crossvalidation run. AC found statistically significantly higher dependency between the data sets than K-means ($p < 0.05$; paired t-test), and the actual values of the log Bayes factors were -23.45 for AC and -48.96 for K-means. This confirmed that at least a subset of the genes has a non-random dependency between TF-binding

Fig. 2. Dendrogram of the hierarchical clustering visualizing the similarities between all the genes clustered with 100 bootstrap AC models. Vertical axis represents the average dissimilarity of the genes: 100 means that a pair of genes never occur in the same significantly dependent cross-cluster in 100 bootstrap runs, and 0 that they always co-occur. Note that there is a mass of genes whose dissimilarity, or co-occurances in the different cross-cluster, is over 80, which was the cutoff threshold to produce the final clusters. Several very reliable clusters can also be seen, as downward protruding peaks.

and expression (discernible from these data sets), although the data globally does not show dependency (log Bayes factor is negative).

After these preliminary checks we used the AC to search for salient dependencies between the stress expression data and the TF binding data. A similarity matrix was produced by bootstrap analysis with 100 samples as described in Sec. 2.5, and summarized by hierarchical clustering. Figure 2 shows a few clear clusters interspersed within a background that shows no apparent dependencies. We cut the dendrogram at the height of 80. This defines a threshold on reliability: if genes occur together, within significant clusters, more than in 20 of the 100 bootstrap AC:s on the average, their association is defined reliable.

We validated the clusters extracted from Fig. 2 by investigating the distribution of earlier-found putative ESR genes within them. Since we had designed the models to hunt for regulation patterns of ESR genes, we expected some of our clusters to consist of ESR genes. Indeed, upregulated ESR genes were enriched statistically significantly in 14 out of the 51 clusters (hypergeometric distribution; p-value < 0.001), and downregulated ESR genes in 12 of them. This confirms that our method has succeeded in capturing stress-related genes in clusters.

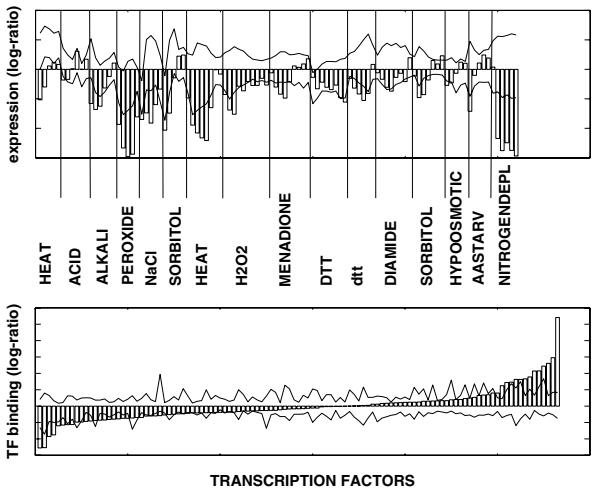

Fig. 3. A gene cluster related to cell cycle, revealing both how cell-cycle machinery is driven down under stress, and the putative regulators for that set of genes. Upper figure represents the mean expression profile (bars) of the genes with their confidence intervals (lines, computed by random sampling) revealing how the genes are downregulated practically in every treatment, and thus conveying information about the shut down of the cell cycle machinery. The lower figure represents the mean TF binding profile (with confidence intervals) of the genes revealing several significant, strong TF bindings. The most interesting of these are analyzed in the text.

3.4. *Biological interpretation*

For more detailed interpretation of the clusters they were analyzed with EASE[11] to find significant enrichments of gene ontology classes. In total we found 14 statistically significant enriched (Bonferroni corrected p-value from Fisher's exact test < 0.05) GO slim classes[‡] in our 51 clusters. Additionally the enrichments of ESR genes as well as interesting, nonrandom TF bindings were used as indicators to select clusters for the analysis. We present here representative examples of the biological interpretations of the clusters.

The cluster in Fig. 3 is an example of a set of genes that are not specifically associated to stress, but obviously behave very homogeneously under stress. The cluster actually contains only two genes

[‡]go_slim_mapping.tab at ftp://genome-ftp.stanford.edu/pub/yeast/data_download/literature_curation/.

known to be ESR genes. Nevertheless, this relatively large cluster can be used to demonstrate two characteristic predictions obtained using AC and confirmed by biological observations. First, this cluster is highly enriched in genes involved in the process of cell cycle (12 out of 57, Bonferroni corrected p-value 3.5e-5). This reflects the coordinated expression of also other genes than the ESR genes, in other words coordination can be highlighted in most experimental conditions. Second, AC will propose a set of transcription factors involved in the regulation of the member genes of the cluster. This is of special value in this case, because although co-ordinated interactions between different signal transduction pathways are essential in biological systems, interpathway connections are difficult to identify. The two most prominent transcription factors of this cluster are coded by SWI4 (YER111C) and FKH2 (YNL068C), which both are known to be involved in cell cycle control. However, they operate on different parts of the process, as shown by Shapira and coworkers for the Forkhead factor.[23]

The significant TF bindings in the same cluster also include ASH1, which is not directly related to the cell cycle process but rather to mating type selection. However, mating-type switching in the yeast is a multi-step programme, which enables Ash1p to asymmetrically localize to the daughter cell nucleus at the end of cell division in order to prevent the daughter cell from switching mating type. Thereby, it is interesting to see that AC has grouped ASH1 together with SWI4 and FKH2.

By far the most consistent cluster is the one in which all genes are down-regulated in every treatment and also classified as downregulated ESR genes.[9] More than 90 of its 100 genes have a GO annotation referring to protein biosynthesis (P-value 1.2e-86). Now, these genes are well known to be strictly co-regulated and therefore this finding is not in itself very special, although it confirms the efficiency of the clustering method. However, a closer look at the associated transcription factors is interesting, especially if one looks at factors such as SFP1 (YLR403W). This factor inhibits nuclear protein localization when present in multiple copies and is thereby a regulator of transcription factor activity. It has been associated to the process of cell size. Yeast establishes this balance by enforcing growth to a critical cell size prior to cell cycle commitment ('Start') in late G1 phase. Interestingly, Jorgensen and collaborators have show that SFP1 is one of two potent negative regulators of 'Start'.[12] SFP1 is shown to be an activator of the ribosomal protein (RP) and ribosome biogenesis (Ribi) regulons, the transcriptional programs that dictate ribosome synthesis rate in accord with environmental and intracellular conditions. This clearly shows that the prediction of associated transcription factors by the AC algorithm has a potential to produce meaningful regulator hypotheses.

Finally, we demonstrate some novel hypotheses obtained by associative clustering. The cluster in Fig. 4 contains 11 genes, of which 10 belong to ESR as defined by Gasch *et al.*[9] These genes contain mainly (7) hypothetical open reading frames (as classified in Stanford Genome Database). Two of them are annotated as responding to stress. Of the four better-known genes two are involved in glutamate and glutatione catabolism and their expression is known to be expressed mainly as a response to nitrogen starvation or oxidative stress, both typical

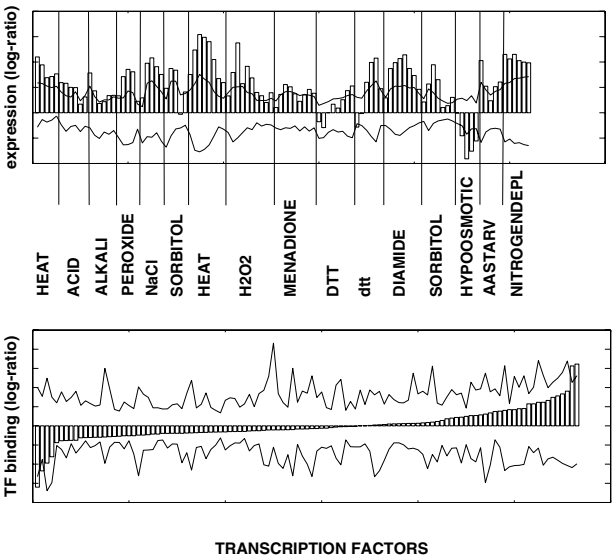

Fig. 4. A gene cluster consisting almost totally of ESR genes, which still are largely unknown. Upper figure represent the mean expression profile of the genes with their confidence intervals (computed by random sampling) revealing how the genes are upregulated practically in every treatment, confirming the earlier definition as upregulating ESR genes. The lower figure represents the mean TF binding profile (with confidence intervals) of the genes revealing two significant, strong TF bindings, which are now potential regulators for these genes.

stress inducers. The associated transcription factor DAL82 (YNL314W) is a positive regulator of allophanate inducible genes. The other highly associated TF STB1 (YNL309W) has to be associated to SWI6 to be activated whereafter it is involved in G1/S transition during the cell cycle.

4. Discussion

We have demonstrated the effectiveness of exploratory dependency modeling for characterizing yeast gene expression under stress, and its regulation.

We applied generalized canonical correlations (gCCA) in a novel way to multiple stress expression sets to produce one representation for the sets, which preserves mutual information between the sets. This preprocessed data was then clustered with AC to maximize its dependency with binding profiles of a set of regulators.

Biological relevance of the clusters was confirmed with several tests and database searches. We can conclude that our approach succeeded notably well both in confirming some known facts, and in generating new hypotheses.

From the technical point of view, the main benefit from gCCA in this work was the reduction of the multiple data sets into one representation, in a mutual information-preserving way. However, the resulting two stage approach is naturally suboptimal, since combining the projection and clustering into one method should improve the results. This is one possible future research direction.

Another interesting future research direction is to use a kernel version of gCCA[24,1] in combining the data sets. In principle, the nonlinear kernel CCA is superior in finding the correlating components when the data is not normally distributed. However, further work will be required to study the regularization of the kernel CCA, the choice of kernel function, its application in the integration of the data sets, and its interpretation. There exist additionally other promising kernel-based data integration methods in bioinformatics.[19]

From a biological perspective, the TF-binding data is problematic since it has been measured in optimal environmental conditions, while the gene expression has been measured under environmental stress. This implies that, for example, the stress regulators such as MSN2p, that are known to bind to genes only under stress, cannot be found in this type of analysis. The results are expected to improve when new binding data measured under stress become available. In spite of the slight deficiency in the data, we managed to extract many biologically meaningful clusters and hypotheses for their regulators. Moreover, this kind of combination of data sets can be seen as a complementary study to previous ones,[5,9] having potential to reveal both totally new stress regulators as well as cell mechanisms that are regulated in concert both under stress and in normal growth conditions.

Acknowledgments

This work has been supported by the Academy of Finland, decisions #79017 and #207467.

References

1. F. R. Bach and M. I. Jordan, Kernel independent component analysis, *J. Machine Learning Research* **3** (2002) 1–48.
2. S. Basu, M. Bilenko and R. J. Mooney, A probabilistic framework for semi-supervised clustering, in *Proc. Tenth ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD-2004)* (2004), pp. 59–68.
3. S. Becker, Mutual information maximization: Models of cortical self-organization, *Network: Computation in Neural Systems* **7** (1996) 7–31.
4. M. A. Beer and S. Tavazoie, Predicting gene expression from sequence, *Cell* **117** (2004) 185–198.
5. H. C. Causton, B. Ren, S. S. Koh, C. T. Harbison, E. Kanin, E. G. Jennings, T. I. Lee, H. L. True, E. S. Lander and R. A. Young, Remodeling of yeast genome expression in response to environmental changes, *Molecular Biology of the Cell* **12** (2001) 323–337.
6. B. Efron and R. Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall, New York, 1993).
7. N. Friedman, Inferring cellular networks using probabilistic graphical models, *Science* **303** (2004) 799–805.
8. N. Friedman, O. Mosenzon, N. Slonim and N. Tishby, Multivariate information bottleneck, in *Proc. UAI'01, The Seventeenth Conf. Uncertainty in Artificial Intelligence* (Morgan Kaufmann Publishers, San Francisco, CA, 2001), pp. 152–161.
9. A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein and P. O. Brown, Genomic expression programs in the response of yeast cells to environmental changes, *Molecular Biology of the Cell* **11** (2000) 4241–4257.
10. I. J. Good, On the application of symmetric Dirichlet distributions and their mixtures to contingency tables, *Annals of Statistics* **4**(6) (1976) 1159–1189.

11. D. A. Hosack, G. Dennis Jr., B. T. Sherman, H. C. Lane and R. A. Lempicki, Identifying biological themes within lists of genes with ease, *Genome Biology* **4**(R70) (2003).
12. P. Jorgensen, I. Rupes, J. R. Sharom, J. R. Broach L. Schneper and M. Tyers, A dynamic transcriptional network communicates growth potential to ribosome synthesis and critical cell size, *Genes Development* (2004). Electronic publication ahead of print, PubMed ID PMID: 15466158.
13. S. Kaski, J. Sinkkonen and A. Klami, Discriminative clustering, *Neurocomputing* (2005), in press.
14. S. Kaski, J. Nikkilä, J. Sinkkonen, L. Lahti, J. Knuuttila and C. Roos, Associative clustering for exploring dependencies between functional genomics data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, in press.
15. M. K. Kerr and G. A. Churchill, Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments, in *Proc. National Academy of Sciences* **98** (2001) 8961–8965.
16. J. R. Kettenring, Canonical analysis of several sets of variables, *Biometrika* **58**(3) (1971) 433–451.
17. R. D. King, Applying inductive logic programming to predicting gene function, *AI Magazine* **25** (2004) 57–68.
18. S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1959).
19. G. R. G. Lanckriet, T. D. Bie, N. Cristianini, M. I. Jordan and W. S. Noble, A statistical framework for genomic data fusion, *Bioinformatics* **20**(16) (2004) 2626–2635.
20. T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Tomphson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford and R. A. Young, Transcriptional regulatory networks in Saccharomyces cerevisiae, *Science* **298** (2002) 799–804.
21. J. Nikkilä, C. Roos and S. Kaski, Exploring dependencies between yeast stress genes and their regulators, in *Proc. Int. Conf. Intelligent Data Engineering and Automated Learning (IDEAL 2004)*, eds. Z. R. Yang, R. Everson and H. Yin (Springer, 2004), pp. 92–98.
22. E. Segal, R. Yelensky and D. Koller, Genome-wide discovery of trancriptional modules from dna sequence and gene expression, *Bioinformatics* **19**(Suppl 1) (2003) 273–282.
23. M. Shapira, E. Segal, M. Shapira and D. Botstein, Disruption of yeast forkhead-associated cell cycle transcription by oidative stress, *Molecular Biology of the Cell* (2004). Electronic publication ahead of print.
24. J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis* (Cambridge University press, 2004).
25. N. Tishby, F. C. Pereira and W. Bialek, The information bottleneck method, in *Proc. 37th Annual Allerton Conf .Communication, Control, and Computing*, eds. B. Hajek and R. S. Sreenivas (Univ. of Illinois, Urbana, 1999), pp. 368–377.
26. K. Wagstaff, C. Cardie, S. Rogers and S. Schroedl, Constrained k-means clustering with background knowledge, in *Proc. of Int. conf. Machine Learning* (2001) 577–584.