Exploring Dependencies between Yeast Stress Genes and their Regulators*

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

[1] Neural Networks Research Centre, Helsinki University of Technology, P.O. Box 5400, FIN-02015 HUT, Finland.
[2] Medicel Oy, Huopalahdentie 24, FIN-00350 Helsinki, Finland.
[3] Dept. of Computer Science, P.O. Box 26, FIN-00014 University of Helsinki, Finland.

Abstract. An environmental stress response gene should, by definition, have common properties in its behavior across different stress treatments. We search for such common properties by models that maximize common variation, and explore potential regulators of the stress response by further maximizing mutual information with transcription factor binding data. A computationally tractable combination of generalized canonical correlations and clustering that searches for dependencies is proposed and shown to find promising sets of genes and their potential regulators.

1 Introduction

Yeast stress response has been studied intensively during recent years [2, 4]. 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, 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 will search for such properties among all the other variation in the gene expression profiles by maximizing the variation that is common to the stress treatment data sets.

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

* This work was supported by the Academy of Finland, decisions 79017 and 207467. We wish to thank Eerika Savia for insights about gCCA.

use so-called *associative clustering* [10] which maximizes dependency between data sets and hence should suit the task perfectly. In practice, a linear projection method scales better to multiple data sets, and we use a generalization of *canonical correlations* as a preprocessing to reduce the number of data sets.

Clusterings that maximize mutual information have been formalized also in the information bottleneck [11] framework. Our clusterings can be viewed as extensions of the current information bottleneck algorithms, suitable for finite sets of continuous-valued data.

Graphical models of dependencies between variables are another related popular formalism. They have been applied to modeling regulation of gene expression [3]. 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 probabilistic models.

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. This fits perfectly the present application.

2 Methods

2.1 Generalized canonical correlation analysis

Canonical correlation analysis (CCA) is a classical method for finding (linear) dependencies between two data sets. It searches for a component from each set, such that the correlation between the components is maximized. The next pair of components maximizes the correlation subject to the constraint of being uncorrelated with the first components. For normally distributed data CCA maximizes mutual information.

There exist several generalizations of CCA to multiple data sets. We selected the generalization (abbreviated gCCA here) that preserves the connection of CCA to mutual information [1]. The gCCA for M multivariate, normally distributed variables $V_1, ..., V_M$ can be formulated as a generalized eigenvalue problem $C\xi = \lambda D\xi$, where $\lambda = 1 + \rho$ (ρ being the canonical correlation), C is the covariance matrix of the concatenated original variables $V = [V_1^T \; ... \; V_M^T]^T$, $\xi = [\xi_1^T \; ... \; \xi_M^T]^T$ are the eigenvectors, and $D = diag(C_{11}, ..., C_{MM})$ is the block diagonal matrix consisting of the covariance matrices of the original variables. For more details see [1] and [8] where the closest method is SUMCOR.

We use gCCA in a possibly novel way for dimensionality reduction. Standardizing the data, i.e. making $C_{mm} = I$, does not affect the eigenvalues, but reduces the eigenvalue equation (for the standardized data) to a standard eigenvalue problem $C\xi = \lambda\xi$ that produces an orthogonal set of ξ. The whole concatenated, standardized data set is then projected onto the reduced-dimensional space formed of these eigenvectors. It can be shown that the first p eigenvectors $\xi^{(1)}, ..., \xi^{(p)}$ (corresponding to the p largest eigenvalues) then preserve maximally the mutual information between the original variables V_m. As a result we obtain a new, low-dimensional representation Y that preserves the mutual dependency between the original variables: $Y = \Xi^T V$, where $\Xi = [\xi^{(1)}, ..., \xi^{(p)}]$.

2.2 Associative clustering

Given a set of paired observations $\{(\mathbf{x}_k, \mathbf{y}_k)\}_k$ from two continuous, vector-valued random variables X and Y, associative clustering (AC) [10] produces two sets of partitions, one for X and the other for Y. The aim is to make the cross-partition contingency table represent as much of the dependencies between X and Y as possible. The Bayesian criterion 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)

$$\frac{P(\{n_{ij}\}|\bar{H})}{P(\{n_{ij}\}|H)} \propto \frac{\prod_{ij} \Gamma(n_{ij} + \alpha_{ij})}{\prod_i \Gamma(n_{\cdot i} + \alpha_i) \prod_j \Gamma(n_{j\cdot} + \alpha_j)} , \tag{1}$$

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, parameterized by a respective prototype vector $\mathbf{m}$. In the optimization the partitions are smoothed to facilitate the application of gradient-based methods. More details in [10].

Note that our use of Bayes factors is different from their traditional use in hypothesis testing, cf. [5]. In AC we do not test any hypotheses but the Bayes factor is maximized to explicitly hunt for dependencies. It may also be worth pointing out that we do not expect to find dependencies between the margins for the *whole data*. Instead, a reproducible dependency between the margin clusters will be regarded as evidence for dependency for some *subsets of the data*. Reproducibility is in this work evaluated by comparing, with cross-validation, contingency tables produced by AC and independent K-means clusterings computed for both spaces separately. Should AC not find a non-random dependency, Bayes factors from AC and K-means should be about equal for test data.

Uncertainty in the AC clusters, stemming from the finiteness of the data sets, is accounted for by a bootstrap analysis [7]. We wish to find clusters (contingency table cells) that signify dependencies between the data sets, and that are reproducible. The AC produces an estimate of how unlikely a cell is given the margins are independent (details in [10]), and the reproducibility is evaluated by the bootstrap. These two criteria will be combined by evaluating, for every gene pair, how likely they are to occur within the same significant cluster. This similarity matrix will finally be summarized by hierarchical clustering.

2.3 GCCA as preprocessing for AC

Having observed for one set of objects (genes) several multivariate variables $V_1, ..., V_M, X$ (stress treatments and TF-binding) forming several data sets, our aim is to find clusters of genes that maximize the mutual information $\mathrm{MI}(V_1, ..., V_M, X)$, 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 $\mathrm{MI}(V_1, ..., V_M, X) = \mathrm{MI}(V_1, ..., V_M) + \mathrm{MI}((V_1, ..., V_M), X)$ we propose a sequential approximation: first approximate $\mathrm{MI}(V_1, ..., V_M)$ by forming the optimal representation $Y(V_1, ..., V_M)$ with gCCA, then maximize $\mathrm{MI}(Y, X)$ 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 only produces a projection of the data.

3 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 [2, 4]. 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 [9], to search for dependencies with expression patterns. In total we ended up in 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.

4 Results

Dependencies between the 16 different stress treatment time series were first extracted with gCCA. The number of gCCA components was chosen to maximize the mutual information in left-out data, in 20-fold crossvalidation. This resulted in 12 components. gCCA was validated by testing the association of the 12 first gCCA components to genes known to be affected by stress, namely the environmental stress genes (ESR) found in [4]. Most of the components showed significant association to ESR genes (Wilcoxon rank sum test; $p < 0.01$), thus validating the use of gCCA.

This 12-dimensional expression data and the TF-binding data were then clustered with AC (with 30x20 cells in contingency table).

To verify that there are dependencies the AC is able to detect, the contingency table produced by AC was compared to the contingency table obtained with independent K-means clusterings in both data spaces, in a 10-fold crossvalidation run. AC found statistically significantly higher dependency between the data sets than K-means ($p < 0.01$; paired t-test). This confirmed that some subset of the genes has a non-random dependency between TF-binding and expression (discernible from these data sets).

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 Section 2.2, and summarized by hierarchical clustering. Figure 1 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, on average more than in 20 of the 100 bootstrap AC:s, their association is reliable.

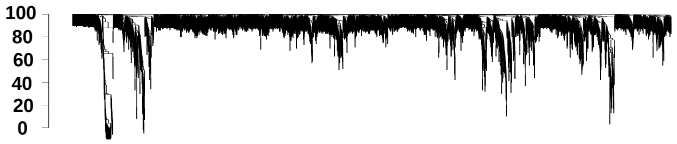

Fig. 1. Hierarchical clustering that summarizes reliable dependencies between stress genes and binding patterns of regulators. The y-axis represents the average distance of the genes in clusters, and each vertical line represents one gene or cluster. The similarity matrix is computed from bootstrap samples of AC. The mass of clusters has an average within-cluster distance between 80 and 100, showing no obvious dependency. The groups of lines protruding downwards from the non-dependent mass are potential reliable gene groups with non-random dependency between their TF-binding and stress-related expression.

We validated the clusters extracted from Figure 1 by investigating the distribution of earlier-found 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.

Finally the clusters were analyzed with EASE [6] to find significant enrichments of gene ontology classes. In total we found 14 statistically significant enriched (Bonferroni corrected p-value < 0.05) GO slim classes[4] in our 51 clusters. The most common (7 out of 14) was the biological process "ribosome biogenesis and assembly" and its subclasses that were found in clusters where genes were strongly downregulated. This is a clear indication of yeast's response to stress

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

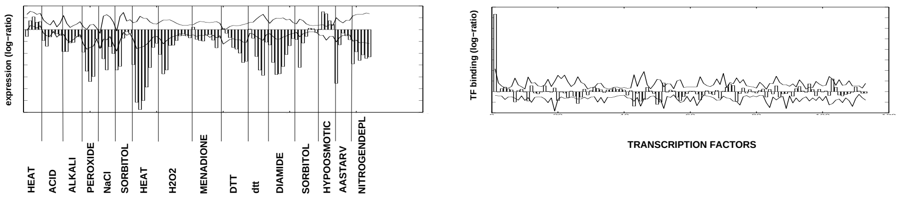

Fig. 2. An example cluster revealing an interesting dependency between gene expression in stress and TF-binding. Left: the average expression profile of the genes in the cluster. Right: the average TF-binding profile of the same genes. Continuous lines depict the 0.999 confidence intervals (computed from 10000 random clusters). The cluster profiles suggest that this set of genes is nearly always downregulated under stress, and that this behaviour might be due to transcription factors with the highest peaks in TF-profile. The specific TFs are currently under analysis.

by driving down the different components of protein synthesis machinery. The TFs in these clusters will be analyzed further; an example is shown in Figure 2.

5 Discussion

We explored dependencies between yeast gene expression under stress and putative gene regulators by dependency maximization methods.

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 the dependency to 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.

References

1. F. R. Bach and M. I. Jordan. Kernel independent component analysis. *JMLR*, 3:1–48, 2002.
2. H. C. Causton et al. Remodeling of yeast genome expression in response to environmental changes. *Molecular Biology of Cell*, 12:323–337, February 2001.
3. N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303:799–805, 2004.
4. A. P. Gasch et al. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11:4241–4257, 2000.
5. I. J. Good. On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Annals of Statistics*, 4(6):1159–1189, 1976.
6. D. A. Hosack et al. Identifying biological themes within lists of genes with ease. *Genome Biology*, 4(R70), 2003.
7. M. K. Kerr and G. A. Churchill. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *PNAS*, 98:8961–8965, 2001.
8. J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.
9. T. I. Lee et al. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298:799–804, 2002.
10. 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.
11. N. Tishby, F C. Pereira, and W. Bialek. The information bottleneck method. In *37th Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377. Urbana, Illinois, 1999.