Associative clustering to find dependencies between expression profiles and transcription factor binding

Janne Nikkilä[1] janne.nikkila@hut.fi

Christophe Roos[2] christophe.roos@helsinki.fi

Janne Sinkkonen[1] janne.sinkkonen@hut.fi

Samuel Kaski[1] samuel.kaski@hut.fi

[1]Helsinki University of Technology
Neural Networks Research Centre
P.O. Box 9800, FIN-02015 HUT, FINLAND

[2]MediCel Ltd.
System Biology Platform Development
Huopalahdenkatu 24, FIN-00350 Helsinki, FINLAND

Keywords. Associative clustering, Bayes factor, gene expression, mutual information, transcription factor

Introduction

A usual assumption in gene expression analysis is that sets of genes are partially co-expressed due to transcriptional co-regulation. Genes regulated coherently should have common features in their promoter region where regulators, *transcription factors*, bind.

We search for groups of yeast *Saccharomyces cerevisiae* genes having both coherent expression patterns and coherent transcription factor binding patterns. Such clusters should contain genes that are most likely co-regulated in the set of expression measurements. The groups will be used to construct hypotheses for further study.

In previous works genes have been clustered based on their expression patterns, and regularities have been sought in the promoters within the clusters (see e.g. [1]). Alternatively, regularities found in the promoters have been correlated with the expression. These sequential approaches have the disadvantage that since the clusters are computed from only one of the data sets, they cannot find optimally the dependencies.

Structured models of the joint distribution of two or more random variables, represented by different data sets, have also been proposed, see e.g. [4]. They introduce clear improvements to mere clustering, and offer extensive selection of post-processing options to extract clusters and infer interactions between variables. The work in this paper is complementary, and meant to serve in particular the first exploratory "looking at" a new data before constructing more detailed structured models.

We cluster genes with *associative clustering (AC)*, which is an extension of the previously published *discriminative clustering* [5]. AC is designed to cluster a paired data set, for example the expression profile of a gene vs. the binding pattern of transcription factors in its promoter region. The end result of associative clustering is two sets of clusters, one for the expression patterns and the other for the transcription factor bindings, and a contingency table representing the dependencies between the clusters. The cluster sets are optimized to maximize dependencies between them, and the dependencies can be visualized and interpreted with the help of the contingency table.

For each yeast gene we have two different kinds of data: expression measured in 300 conditions [2] and the experimental binding capacity of 113 transcription factors in its promoter region [3]. Finding the set of common genes represented in both data sets resulted in a final data set of 6185 genes. Missing values were imputed by genewise averages in both data sets, and the variances of each dimension were normalized to unity.

Associative clustering

Given a set of paired data samples $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x} \in \mathbb{R}^\mathbf{n}, \mathbf{y} \in \mathbb{R}^\mathbf{m}$ the task of associative clustering is to cluster or partition each data space such that the clusters are maximally dependent. Intuitively speaking, associative clustering prefers partitionings where the data points in one cluster of $\mathbf{x}$ space belong to one partition also in the $\mathbf{y}$ space.

This work has been supportted by the Academy of Finland, grant 50061

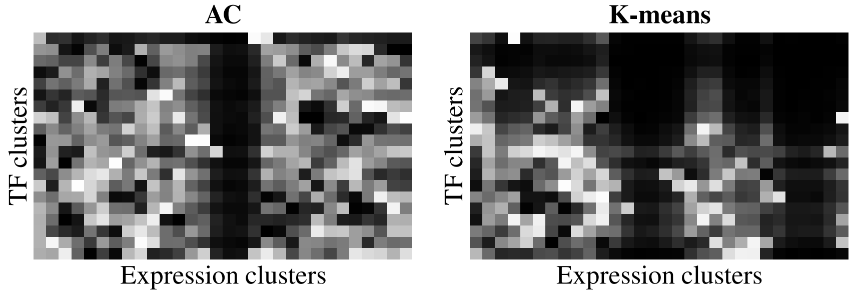

Figure 1. Comparison of contingency tables from associative clustering and from K-means clusterings with leave-out data. The light shades depict expectionally dependent groups of genes and examples of their interpretation are given in Results. It is evident that far more cells in the AC table are exceptionally dependent than in the K-means table. Locations in the tables are not comparable.

AC is computed to maximize the logarithmic Bayes factor (finite data mutual information; technically the Bayes factor between the hypotheses of dependent vs. independent margins in the contingency table)

$$\log BF = \sum_{ji} \log \Gamma(n_{ji} + n^0) - \lambda^{(y)} \sum_{j} \log \Gamma(n_{j\cdot} + n^0) - \lambda^{(x)} \sum_{i} \log \Gamma(n_{\cdot i} + n^0) + \mathit{const.}$$

Here n_{ji} is the number of data pairs where $\mathbf{x}$ belongs to the cluster i and $\mathbf{y}$ to cluster j, $n_{j\cdot} = \sum_i n_{ji}$, $n_{\cdot i} = \sum_j n_{ji}$, n^0 comes from an uninformative Dirichlet prior, and λ:s are regularization parameters chosen using a validation set. The Bayes factor is optimized with respect to the model vectors representing the cluster centroids using the conjugate-gradient method, analogously to DC in [5]. The cluster assignments are smoothed for the optimization.

Results

We first studied whether there are significant dependencies between these specific data sets; it is not self-evident because of all the noise in the data. The contingency table formed by AC was compared with the contingency table produced from two separate K-means clusterings of the two data sets. The comparison was done by ten-fold crossvalidation, in which the Bayes factor was computed from test data for each method for each fold. A paired t-test confirmed that the difference in favour of AC is statistically significant, and thus also confirmed that the dependency between expression and transcription factor binding is statistically significant in these data sets.

Sample contingency tables of AC and K-means have been visualized in Figure 1. The gray shades show the contributions of the table cells to the Bayes factor, the most informative cells being white. We chose a set of the most informative cells and analyzed the genes mapped into them. A sample cluster contained genes known to be related to membrane synthesis and secretion activity, to mitosis and to cytokinesis. These processes occur sequentially, and it is plausible that proteins needed in the successive stages could be produced simultaneously. Hence, the corresponding genes could be regulated by the same transcription factors, which is indeed suggested by the clustering. In another sample cluster a group of genes related to building of the cell wall and some cytoskeletal components, and also to mitochondrial biogenesis were identified. Those clusters contained additionally other genes, many with unknown biological function. The hypothesis to be studied later is that these genes might also be regulated in the same way and contribute to the same biological processes.

References

[1] A. Brazma and J. Vilo. Gene expression data analysis. *FEBS Letters*, 480:17–24, 2000.

[2] T. R. Hughes et al. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.

[3] T. Lee et al. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298:799–804, 2002.

[4] E. Segal, B. Taskar, A. Gasch, N. Friedman, and D. Koller. Rich probabilistic models for gene expression. *Bioinformatics*, 17(Suppl 1):243–252, 2003.

[5] J. Sinkkonen, S. Kaski, and J. Nikkilä. Discriminative clustering: Optimal contingency tables by learning metrics. In *Proc of the ECML'02, 13th European Conference on Machine Learning*, Lecture Notes in Artificial Intelligence 2430, pages 418–430, Berlin, 2002. Springer.