Semi-supervised Clustering of Yeast Gene Expression Data

A. Schönhuth, I.G. Costa, and A. Schliep

Abstract To identify modules of interacting molecules often gene expression is analyzed with clustering methods. Constrained or semi-supervised clustering provides a framework to augment the primary, gene expression data with secondary data, to arrive at biological meaningful clusters. Here, we present an approach using constrained clustering and present favorable results on a biological dataset of gene expression time-courses in Yeast together with predicted transcription factor binding site information.

1 Introduction

Life on the biochemical level is driven by large molecules acting in concert following complex patterns in response to internal and external signals. Understanding these mechanisms has been the core question of molecular biology for the time since discovery of the DNA double helix. Ideally, one would like to identify detailed pathways of interaction. Unfortunately, this is often impossible due to data quality and the superposition of many such pathways in living cells. This dilemma led to the study of modules – sets of interacting molecules in one pathway – as identifying such modules is comparatively easy. In fact, clustering easily available mass data such as gene expression levels, which can be measured with DNA microarrays simultaneously for many genes is one approach for identifying at least parts of modules: for example co-regulated genes which show similar expression levels under several experimental conditions due to similarities in regulation.

The effectiveness of this approach is limited as we cluster based on observable quantities, the gene expression levels, disregarding whether the observed level can arise due to the same regulatory mechanism or not. Considering this information during the clustering should yield biologically more helpful clusters. Here we are

A. Schönhuth(✉)
ZAIK, Universität zu Köln, 50931 Cologne, Germany, E-mail: asa86@cs.sfu.ca

A. Okada et al. (eds.), *Cooperation in Classification and Data Analysis*,
Studies in Classification, Data Analysis, and Knowledge Organizations.
DOI: 10.1007/978-3-642-00668-5_16,

dealing with primary data, the gene expression levels, augmented with secondary data, for example transcription factor (TF) binding information.[1] Unfortunately, such secondary data is often scarce, in particular if we require high quality data.

Constrained clustering constitutes a natural framework. It is one of the methods exploring the gamut from unsupervised to supervised learning and it uses the secondary data to essentially provide labels for a subset of the primary data. Semi-supervised techniques have successfully been employed in image recognition and text classification (Lange et al. 2005; Lu and Leen 2005; Nigam et al. 2000). Hard constraints for mixture models (Schliep et al. 2004) were, to the best of our knowledge, the first application of constrained clustering in bioinformatics which showed the effectiveness of highest quality *must-link* or positive constraints indicating pairs of genes which should be grouped together. Here we use a soft version (Lange et al. 2005) which can cope with positive (*must-link*) and negative constraints (*must-not-link*) which are weighted with weights from [0, 1].

Constrained learning is used to estimate a mixture model where components are multi-variate Gaussians with diagonal covariance matrices representing gene expression time-courses. The secondary data consists of occurrences of transcription factor binding sites in upstream regions of yeast genes. Its computation is based on methods proposed in Rahmann et al. (2003) and Beer et al. (2004). The more transcription factor binding sites (TFBS) two yeast genes have in common, the more likely it is that they are regulated in a similar manner, which is reflected in a large positive constraint. Previously, we showed that even modest noise in the data used for building constraints actually will result in worse clustering solutions (Costa and Schliep 2006); the main contribution here is the careful construction of the secondary dataset and the method for evaluating the effectiveness of using constraints.

2 Methods

A mixture model (McLachlan and Peel 2000) is defined as

$$\mathbf{P}[x_i|\Theta] = \sum_{k=1}^{K} \alpha_k \mathbf{P}[x_i|\theta_k], \tag{1}$$

where $X = \{x_i\}_{i=1}^{N}$ is the set of (observed) data. The overall model parameters $\Theta = (\alpha_1, \ldots, \alpha_K, \theta_1, \ldots, \theta_K)$ are divided into the probabilities $\alpha_k, i = 1, \ldots, K$ which add to unity for the model components $\mathbf{P}[x_i|\theta_k]$ and the $\theta_k, k = 1, \ldots, K$, which describe the multi-variate Gaussians components of the mixture. One now aims at maximizing (1) by choosing an optimal parameter set Θ. This problem is routinely solved by the EM algorithm, which finds a local optimum for the above

[1] Transcription factors are essential for inhibiting or enhancing the production of proteins encoded in a gene.

function by involving a set of hidden labels $Y = \{y_i\}_{i=1}^{N}$, where $y_i \in \{1, \ldots, K\}$ is the component, which generates data point x_i. For details of the EM algorithm see Bilmes (1998).

In addition to the data x_i one is now given a set of positive respectively negative constraints w_{ij}^{+} resp. $w_{ij}^{-} \in [0, 1]$, which reflect the degree of linking of a pair of data points $x_i, x_j, 1 \leq i < j \leq N$. The task is to integrate these constraints meaningfully and consistently into the EM routine. We will explain the essence of the solution proposed in Lange et al. (2005) and applied in Lu and Leen (2005) and Costa and Schliep (2006). Computation of the Q-function in each step of the EM-algorithm requires the computation of the posterior distribution $P[Y|X, \Theta]$ over the hidden labels y_i, where Θ is an actual guess for the parameters. By Bayes' rule we have

$$\mathbf{P}[Y|X, \Theta] = \frac{1}{Z} \cdot \mathbf{P}[X|Y, \Theta] \cdot \mathbf{P}[Y|\Theta], \tag{2}$$

where Z is a normalizing constant. The constraints are now incorporated by, loosely speaking, choosing as prior distribution $\mathbf{P}[Y|\Theta]$ the one, which is "most random" without that the constraints and that the prior probabilities α_k in Θ get violated. In other words, we choose the distribution, which obeys the *maximum entropy* principle and is called the *Gibbs* distribution (see Lange et al. 2005 for a theoretical setting and Lu and Leen 2005 for formulas and further details):

$$\mathbf{P}[Y|\Theta] = \frac{1}{Z} \prod_i \alpha_{y_i} \prod_{i,j} \exp[-\lambda^{+} w_{ij}^{+}(1 - \delta_{y_i y_j}) - \lambda^{-} w_{ij}^{-} \delta_{y_i y_j}], \tag{3}$$

where Z is a normalizing constant. The Lagrange parameters λ^{+} and λ^{-} define the penalty weights of positive and negative constraints violations. This means that increasing λ^{+}, λ^{-} leads to an estimation, which is more restrictive with respect to the constraints. Note that computing (2) is usually infeasible and thus requires a *mean field approximation* (see again Lange et al. 2005 and Lu and Leen 2005 for details). Note, finally, that when there is no overlap in the annotations – more exactly, $w_{ij}^{+} \in \{0, 1\}$, $w_{ij}^{-} \in \{0, 1\}$, $w_{ij}^{+} w_{ij}^{-} = 0$, and $\lambda^{+} = \lambda^{-} \sim \infty$ – we obtain hard constraints as the ones used in Schliep et al. (2005), or as implicitly performed in Pan (2006).

2.1 The Gene-TFBS-Matrix

The computational basis for the constraints is a binary valued incidence matrix, where the rows correspond to genes and the columns correspond to transcription factor binding sites (TFBS). A one indicates that, very likely, the TFBS in question occurs in the upstream region of the respective gene.

In a first step TFBS profiles were retrieved from the databases SCPD[2] and TRANSFAC.[3] In addition to consensus sequences for reported profiles we computed conserved elements in the upstream regions of the yeast's genes by means of the pattern hunter tool AlignACE.[4] In a second step we removed redundant patterns resulting in 666 putative TFBS sequence patterns. We then computed positional weight matrices (PWM) from these patterns by using G-C-rich background frequencies to contrast the patterns, following Rahmann et al. (2003).

With the PWMs we computed p-values for the occurrence of a TFBS in the upstream region of a gene by means of the following Monte Carlo approach. First, we generated 1,000 G-C-rich sequences of the length of the upstream sequences ($800bp$). We then computed a score for each of the 1,000 random sequences and each of the 666 PWMs by sliding a window of the length of the PWM in question over the sequence and adding up the values given by the PWM. We thus obtained, for each of the PWMs, a distribution of scores in sequences of length 800. We finally set a one in the Gene-TFBS-Matrix (GT-matrix) if the score of an upstream sequence of a gene (obtained by the same procedure as for the random sequences) was below a p-value of 0.001 compared to the distribution given through the random sequences. We note that we chose a very restrictive p-value as TFBS analysis usually is very easily corrupted by false positive hits (Rahmann et al. 2003; Claverie and Audic 1996) and false positives negate the benefits of constrained clustering.

2.2 Constraints

From the `GT-Matrix` we compute positive and negative constraints. We remind the reader that, by means of the `GT-Matrix` we have, for each of the genes, a binary valued vector of length 666. One is now tempted to, say, define the positive constraint between two genes to be proportional to the number of positions where the binary vectors of the two genes have a one in common (thus indicating that there is a transcription factor acting on both of the genes) and, likewise, to set the negative constraint to be proportional to the number of positions where exactly one of the genes has a one (thus indicating that there is a transcription factor which acts on one but not on both of the genes). Yet, although we expect seeing a one in only one of 1,000 genes in each of the columns of the matrix according to the p-value of 0.001, there are PWMs, which occur frequently (up to 90%) in the genes' upstream sequences. This indicates that there are heterogeneities in the upstream regions in general. It may also be due to the computation of the TFBSs as conserved elements of the upstream sequences themselves.

To address this we computed for each TFBS z the frequency of occurrence p_z within the genes and defined the positive (w_{ij}^{+}) and negative (w_{ij}^{-}) constraints for

[2] Saccharomyces cerevisiae promoter database, http://cgsigma.cshl.org/jian.

[3] The transcription factor database, http://www.gene-regulation.de.

[4] Motif finding algorithm, http://atlas.med.harvard.edu.

two genes i and j as follows. Let M_{iz} denote the `GT-Matrix` entry for gene i and TFBS z and set

$$w_{ij}^{+} := \gamma^{+} \cdot \#\{x \ : \ p_z^2 \leq 0.01,\ M_{iz} = M_{jz} = 1\}.$$

That is, w_{ij}^{+} is up to a scaling factor γ^{+}, the number of TFBSs, which occur with a p-value of 0.01 or less in both genes i and j. Similarly, we define

$$w_{ij}^{-} := \gamma^{-} \cdot (\#\{z \ : \ p_z(1-p_z) \leq 0.01,\ M_{iz} = 1, M_{jz} = 0\} + \#\{z \ : \ p_z(1-p_z) \leq 0.01,\ M_{iz} = 0, M_{jz} = 1\}).$$

2.3 Relevant Constraints

Constrained clustering profits from information of two datasets – the original, primary dataset and the secondary one, from which constraints are computed. When the influence of the secondary dataset is increased, cluster results change. To identify which constraints cause changes we computed the pairs of genes in one cluster, which were in the same cluster in the unconstrained clustering and in distinct clusters in the constrained case or vice versa. Lists of positive and negative constraints for pairs of genes identified ranked by constraint weight serve as the basis for further analysis. This way we identified the TFBSs which had the largest contribution to changes in the clustering.

3 Results

As in Costa and Schliep (2006) we used 384 yeast cell cycle gene expression profiles (YC5) for analysis. YC5 is one of the rare examples of a dataset where high quality labels are available for each gene as each of them is assigned to one of the five mitotic cell cycle phases. Because of the synchronicity of the profiles within one group (corresponding to one of the five phases), we opted for multivariate Gaussians with diagonal covariance matrices as components in the mixture model. We initialized the mixture estimation procedures by means of an initial model collection algorithm presented in Schliep et al. (2005). The clustering solution was obtained from the mixture by assigning each data point to the component of highest posterior probability.

3.1 Clustering Statistics

We estimated mixtures for varying values of the Lagrangian parameters λ^{+}, λ^{-}. Let TP resp. TN denote the amounts of pairs of genes correctly assigned to one resp.

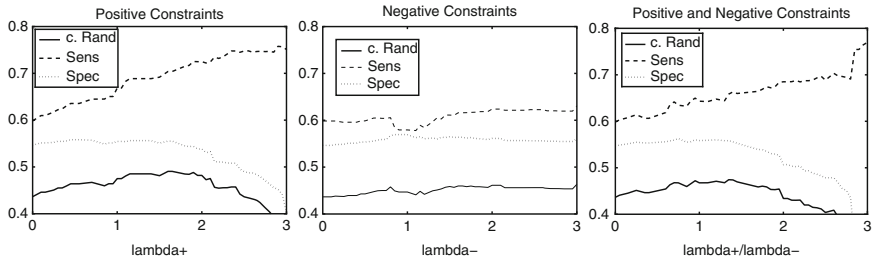

Fig. 1 We depict the CR, Spec and Sens with only positive (*left*), only negative (*middle*) and both positive and negative (*right*) constraints for increasing values of the Lagrangian parameters λ^+, λ^-

two clusters out of P resp. N many according to the true labels. Then, we computed Sens $= \frac{TP}{P}$ and Spec $= \frac{TN}{N}$, and the corrected Rand, which can be perceived as a significance level for the clustering of being distinct from a random distribution of the genes over the clusters, to monitor the effects of an increasing influence of the constraints (Fig. 1).

While the positive constraints improve sensitivity, the negative constraints slightly improve specificity. One also sees a considerable improvement of the corrected Rand for the addition of positive constraints and a slight improvement for the negative constraints. Taking into account both positive and negative constraints one sees improvements in all of the three statistics. However, there does not seem to be a synergy of the positive effects of the two kinds of constraints. This may be an indication for contradictions within the constraints and suggests some "contradiction purging" as a future area of research.

3.2 Gene Ontology Statistics

To validate the clustering quality from a biological point of view we compare the p-values from enrichment of Gene Ontology (GO) terms in a procedure similar to the one performed in Ernst et al. (2005). More specifically, we computed GO term enrichment using GOStat (Beissbarth and Speed 2004) for an unconstrained and a constrained ($\lambda^+ = \lambda^- = 1.35$) mixture estimation as described above. We selected all GO terms with a p-value lower then 0.05 in both clusterings and plotted the $-\log(p\text{-values})$ of these terms in Fig. 2.

We found smaller p-values for the constrained clustering and compile a list of GO Terms, which display high log-ratios in Table 1. The constrained case had 16 of such GO terms, 10 out of these are directly related to biological functions or cell compartments related to cell cycle (big dots[5] in Fig. 2 and GO terms in italic in Table 1). On the other hand, only five GO terms had a higher enrichment in the unconstrained case, all with a significant lower log ratio then in the constrained case.

[5] Due to overlap, only eight big dots are visible.

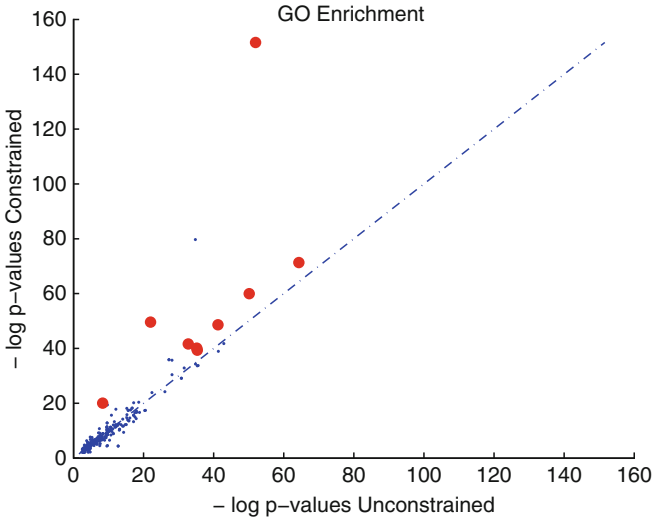

Fig. 2 Scatter plot comparing the GO Term enrichment of the unconstrained (x-axis) and constrained (y-axis) results. Points above the *diagonal line* indicate higher enrichment in the constrained case, while values below indicate higher enrichment in the unconstrained case

From those, the first four are related to chromatin structure and nucleosome, which is related to the S phase of cell cycle.

As described in Sect. 2.3 we computed the constraints which had a relevant impact on the clustering statistics. We found (not shown) that the TFBS data particularly helped correctly classifying genes, which belong to cell cycle phases *late G1* and *S* which is consistent with the gene expression time-course dataset used. Further manual analysis of the relevant constraints and investigation of the TFBSs involved will likely provide insights in mechanisms which are not discoverable from gene expression alone.

4 Conclusion

Constrained clustering is a very useful tool for analyzing heterogeneous data in molecular biology, as there is often an abundant primary data source available (e.g., gene expression, sequence data) which can be made much more useful by integration of high-quality secondary data. However, as the results by Costa and Schliep (2006) show, constrained clustering cannot be applied straight-forwardly even to secondary data sources which are routinely used for biological validation of clustering solutions. Point in case: the *predicted* TFBS information used here improves results whereas the *experimental* chip-on-chip data used by Costa and Schliep (2006) does not. This is likely due to higher error rates in the experimental data and a lack of quality measure for each individual experiment, which precludes filtering

Table 1 List of GO Terms for which the log ratio of the p-values is higher then 4.0 (or $|\log((p\text{-values const.})/(p\text{-values unconst.}))| > 4.0$). Positive ratios indicate a higher relevance of the term in a cluster from the constrained case, while negative ratios indicates higher relevance in a cluster from the unconstrained case

GO Term ID	GO Term	p-value log ratio
GO:0005694	*Chromosome*	99.6581
GO:0009719	Response to endogenous stimulus	44.9090
GO:0000278	*Mitotic cell cycle*	27.6137
GO:0003677	*DNA binding*	11.7053
GO:0044427	*Chromosomal part*	9.7880
GO:0007010	Cytoskeleton organization and biogenesis	9.7352
GO:0000228	*Nuclear chromosome*	8.9036
GO:0043232	Intracellular non-membrane-bound organelle	8.6498
GO:0043228	Non-membrane-bound organelle	8.6498
GO:0044454	*Nuclear chromosome part*	7.5673
GO:0007049	*Cell cycle*	7.4107
GO:0006259	*DNA metabolism*	6.9792
GO:0044450	Microtubule organizing center part	5.6984
GO:0006281	*DNA repair*	4.9234
GO:0007017	Microtubule-based process	4.7946
GO:0006974	*Response to DNA damage stimulus*	4.0385
GO:0000786	Nucleosome	−8.3653
GO:0000788	Nuclear nucleosome	−8.3653
GO:0000790	Nuclear chromatin	−5.1417
GO:0000785	Chromatin	−5.1333
GO:0016043	Cell organization and biogenesis	−4.8856

on quality. Noise reduction in constraints, resolution of conflicts between positive and negatives constraints and measure of constraint relevance are open questions which need to be addressed.

References

BEER, M. A. and TAVAZOIE, S. (2004). Predicting gene expression from sequence. *Molecular Biology of the Cell, 117*, 185–198.

BEISSBARTH, T. and SPEED, T. P. (2004): GOStat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics, 20:*1464–1465.

BILMES, J. A. (1998): A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *Technical Report TR-97-021, International Computer Science Institute, Berkeley, CA*.

CLAVERIE, J. M. and AUDIC, S. (1996): The statistical significance of nucleotide position-weight matrix matches. *CABIOS, 12(5):*431–439.

COSTA, I. and SCHLIEP, A. (2006): On the feasibility of heterogeneous analysis of large scale biological data. In *ECML/PKDD Workshop on Data and Text Mining for Integrative Biology, pages* 55–60.

ERNST, J., NAU, G. J. and BAR-JOSEPH, Z. (2005): Clustering short time series gene expression data. *Bioinformatics, 21:* i159–i168.

LANGE, T., LAW, M. H. C., JAIN, A. K. and BUHMANN, J. M. (2005): Learning with constrained and unlabelled data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 1, pages* 731–738.

LU, Z. and LEEN, T (2005): Semi-supervised learning with penalized probabilistic clustering. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17, pages 849–856*. MIT, Massachusetts.

MCLACHLAN, G. and PEEL, D. (2000): *Finite Mixture Models.* Wiley Series in Probability and Statistics. Wiley, New York.

NIGAM, K., McCALLUM, A. K., THRUN, S. and MITCHELL, T. (2000): Text classification from labeled and unlabeled documents using EM. *Machine Learning, 39(2/3):*103–134.

PAN, W. (2006): Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics, 22(7):*795–801.

RAHMANN, S., MUELLER, T. and VINGRON, M. (2003): On the power of profiles for transcription factor binding site detection. *Statistical Applications in Genetics and Molecular Biology, 2(1):*7.

SCHLIEP, A., STEINHOFF, C. and SCHÖNHUTH, A. (2004): Robust inference of groups of genes using mixtures of HMMs. *Bioinformatics, 20(suppl 1):*i283-i289.

SCHLIEP, A., COSTA, I. G., STEINHOFF, C. and SCHÖNHUTH, A. (2005): Analyzing gene expression time-courses. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2(3):*179–193.