Comparing series of gene groupings with a motif-based heuristic

Patrick Glenisson
Electrical Eng. Dept./SCD
KULeuven
Kasteelpark 10
B-3001 Leuven, Belgium

Janick Mathys
Electrical Eng. Dept./SCD
KULeuven
Kasteelpark 10
B-3001 Leuven, Belgium

Yves Moreau
Electrical Eng. Dept./SCD
KULeuven
Kasteelpark 10
B-300 Leuven, Belgium

Bart De Moor
Electrical Eng. Dept./SCD
KULeuven
Kasteelpark 10
B-3001 Leuven, Belgium

Abstract

Background As biological validations of series of possible gene groupings originating from various cluster algorithms or parameterizations are costly, we develop of a scoring mechanism that provides a one-shot estimate of the relevance of a given gene grouping in terms of upstream regulatory signals present in the derived clusters.

Results Starting from a set of biological assumptions we propose a simple motif-based heuristic to quickly eliminate clustering solutions that yield poor overall results in terms of significant motifs. We test our score both on synthetic and real data, and apply it (1) to determine the optimal number of clusters k in case of hierarchical and k-means clustering, and (2) to compare the solutions of those algorithms. Applied on gold standard yeast expression data, the score suggests an optimal number of clusters, k, that is biologically consistent. Additionally, the score is used to hint at which cluster algorithm is more appropriate to use in a given case.

Conclusions Although the biological assumptions that underpin the heuristic constitute a simplification of biological reality, the score alleviates the problem of exhaustively validating a prohibitive amount of results stemming from either different clustering parameterizations, or various clustering methodologies. It targets a facilitated exploration of good clustering solutions in 'motif space', which proves particularly useful when experimenting with novel clustering strategies, such as those integrating various other information.

1 Background

As clustering gene expression data constitutes an 'ill-posed' problem in the sense that definite objectives are often hard to define, labor-intensive biological evaluations usually start from educated guesses on good clustering parameterizations. Especially when experimenting in silico with various methodologies or parameterizations, quantitative methods for cluster validation can be of great help in choosing 'good' solutions. A wide range of techniques, each formulating the optimality principle differently, already exist to validate genome-wide clusterings. On the one hand, *data*-based scores such as the Figure Of Merit [1], the Rand index [2], the Silhouette-coefficient [3][4], the Gap-statistic [5], or the local stability-based method [6], estimate good solutions based on the statistical properties of the clustered data. However, these class of scores suffer from the drawback that validation is performed on the same data that produced the clusters, without taking into account biological constraints. On the other hand, an increasing amount of tools are becoming available to link cluster results to information from other sources and compare them [7] [8] [9][10].

Over the last years there has been a great interest in detecting regulatory signals (or motifs) in upstream regions of coexpressed genes [11][12][9][13]. However, exploring multiple clustering results over various parameterizations involves a lot of overhead in assessing biological relevance to each of the results (for example, see [14], where authors went to great lengths to validate their clustering method in terms of regulation patterns). We thus observe that (1) purely data-based scores do not necessarily correlate well with clustering solutions that group motifs in a consistent way, and (2) most current ways to evaluate gene groupings in terms of motifs are limited to

manual investigation of statistically enriched clusters. To the best of our knowledge, none of them provide a one-shot estimate of the relevance of all patterns found in the upstream promotor regions in a clustering solution. This supported our motivation for the development of a *motif*-based heuristic to economize on biological validations when the parameter-space is prohibitively large. In this work we present our motif-inspired cluster validation score, abbreviated as M-score, for quickly estimating the validity of multiple clustering runs in terms of significant upstream sequence patterns. For similar reasons of validation overhead, we decided to check our score biologically by focusing on a well-understood data set, for which we have accumulated support from both published results and a yeast biologist.

In Section 2.1 we treat several ways to obtain p-values that describe the statistical overrepresentation of motifs in a clustered data set. We state our biological assumptions and propose the derived M-score in Section 2.2. In Section 3, we first test the score, along with some variations, on synthetic data. Subsequently, we demonstrate its application in determining the optimal number of clusters (in terms of motif enrichment of the solution) using two cluster algorithms on cell-cycle expression data. We complement the discussion of the score's performance with a biological evaluation.

2 Method

The proposed M-score attached to an overall clustering result starts from values describing the significance of a set of defined (or found) motifs in each of the clusters. In this Method section, we therefore first describe how these significance values are computed, followed by a description of the M-score itself.

2.1 Statistical overrepresentation of known motifs

Starting from a predefined set of M Motifs and a partitioning of a data set into k clusters, we define a cluster-by-motif matrix A of size $k \times M$. Each entry contains a p-value that expresses the statistical enrichment of genes within a cluster that contain the given motif within 800bp upstream of the ORF. To compute the p-values in each entry of A, we define a 2×2 contigency table $\{C(i,j)\}_{i=1,2;j=1,2}$, as in Table 1, where r is the number of genes containing the motif, n the cluster size, m the total number of genes that have one or more copies of the given motif upstream, and N the total number of genes in the data set. We use the information captured in the contigency table to derive three options for one-sided p-values that test the hypothesis whether the motif enrichment in the given cluster is higher than what is expected by chance:

Binomial approximation to the hypergeometric distribution When sampling without replacement, the significance of observing at least r genes whose upstream region contain one or more copies of a given motif sequence within a cluster of n genes is given by

$$1 - \sum_{l=1}^{r} \mathcal{H}(m, n, N) \tag{1}$$

.

As this quantity is hard to compute for $r \approx 50$, we use the binomial approximation of the hypergeometric which holds when $N >> n$. The significance then becomes

$$p\text{-value} = 1 - \sum_{l=1}^{r} \mathcal{B}(l, n, \frac{m}{N}). \tag{2}$$

We note that the background probability, $\frac{m}{N}$, is data-dependent since it involves only those genes that remain, for example, after filtering genes with low variance from expression data. This method is used on a regular basis in the literature (e.g., [14])

Pearson χ^2 test for independence The Pearson χ^2 statistic for contigency table C (Table 1) is defined as

$$G^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(c_{ij} - \frac{c_{i.}c_{.j}}{N})^2}{\frac{c_{i.}c_{.j}}{N}} \tag{3}$$

where $c_{\cdot j}$ and $c_{i\cdot}$ denote row- and columnsums respectively. This statistic measures the independence of the random variables *motif occurrence* and *cluster membership*. G^2 is χ^2-distributed with $(2-1)(2-1)$ degrees of freedom and we assume that large values indicate an overrepresentation of the given motif in a certain cluster with at the one-sided significance level

$$p\text{-value} = 1 - \chi^2(1) \tag{4}$$

The distributional properties of Eq.3 are based on the assumption that underlying sampling model is multinomial [15]. Accordingly, in the numerical simulations we conducted, we found the Pearson χ^2 test-statistic less stringent than when modelling explicitly the binomial (Eq.2).

Fisher Exact test for small sample sizes When any of the entries in C become smaller than 8, we use the more appropriate Fisher Exact test, which is more robust to small sample sizes [16]. The hypergeometric probability for C is written as

$$p\text{-value} = \frac{\prod_{i=1}^{2} c_{i\cdot}! \prod_{j=1}^{2} c_{\cdot j}!}{N! \prod_{i,j=1}^{2} c_{ij}!} \tag{5}$$

The one-sided version of Fisher's method requires the construction of the set of all possible 2×2 contingency tables, $\{D^l\}_l$, each table having elements d^l_{ij} $(i, j = 1, 2)$, that are mutually different and satisfy the condition

$$|\frac{d^l_{11}}{d^l_{11} + d^l_{12}} - \frac{d^l_{21}}{d^l_{21} + d^l_{22}}| \geq |\frac{c_{11}}{c_{11} + c_{12}} - \frac{c_{21}}{c_{21} + c_{22}}|.$$

For all contingency tables $\{D^l\}$, we derive p-values as in Eq. 5, and sum them to obtain a final p-value.

As the significance of each motif in is tested along multiple clusters simultaneously, we correct each of the p-value entries in A for multiple testing with the Bonferroni method,

$$p\text{-value}^* = \min(p\text{-value} \cdot k, 1).$$

2.2 M-score

In this section we define the M-score that uses the p-value information contained in a cluster-by-motif matrix A, and measures the amount of biological evidence present in one single clustering run.

Before we proceed, we formulate the three biological assumptions it is built on. Given a cluster-by-motif matrix A as defined in Section 2.1, we assume that

- a motif is less interesting when it (significantly) occurs in many clusters
- provided the set of M motifs is large enough, a cluster that contains a large proportion of the motifs is less likely to be biologically relevant.
- a 'too large' number of clusters is less likely to reflect the true biological diversity underlying the experiment.

The proposed heuristic balances between these three criteria and is composed as follows:

$$\text{M-score} = \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{M} \log(\frac{M}{f_{\{1..M\} \in i}}) \log(\frac{k}{f_{\{1..k\} \ni j}}) \cdot A(i, m) \tag{6}$$

where $f_{\{1..M\} \in i}$ is the number of (significantly) found motifs contained in cluster i, $f_{\{1..k\} \ni j}$ is the (significant) occurrences of motif j over all k clusters, and $A(i, j)$ the p-value for the motif j in cluster i.

The term $\log(\frac{M}{f_{\{1..M\} \in i}})$ can be seen as an *inverse motif frequency*, $\log(\frac{k}{f_{\{1..k\} \ni j}})$ as an *inverse cluster frequency*. They balance between groupings with clusters containing too much significant motifs (typically if that cluster is too big) on the one hand, and groupings in which motifs are too much distributed over all clusters (typically if clusters are too small) on the other hand.

3 Results

We test the motif score from Eq.6 both on synthetic and real data of Saccharomyces cerevisiae. Rather than to exhaustively examine results over multiple data sets or cluster methodologies, our purpose in this work is to demonstrate and validate the score's properties. To this end we select a set of 42 cell-cycle motifs which are listed in the additional data file. Consensus sequences for these motifs were extracted from the TRANSFAC Database [17], and string searches for each motif sequence over all 800bp upstream regions results in a gene-by-motif matrix containing raw counts. Although much more advanced methods than string-based searches for motif detection are available (e.g., [18][9]), this choice still constitutes a reasonable approach (e.g., [12][14]) to validate our setup.

3.1 Synthetic Data

As a proof of principle we select the following set of six motifs : Ace2, ECB, SCB, ICRE, Met31_32p, Rap1 and SCB. We collect all 358 genes carrying either of these motifs in their upstream region and divide them into six groups as in Table 2. Each motif j lies at the basis of a separate group $(1, .., 6)$, with exception of the SCB-motif whose genes are divided over two groups. Each gene g is assigned the artificial expression profile $\{e_{g,j}\}_{j=1..6} = \delta_{(j,group_nr)} + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma)$.

We cluster these 358 genes hierarchically and consider all possible cutoffs for k in the range 2..40. Figure 1 plots the influence of the various motif p-values from Section 2.1 on the M-score in case of $\sigma = 0.2$.

We see that the binomial method results in a peak at the encoded number of clusters, $k_{real} = 6$. According to expectations, computations based on Pearson χ^2-based values behave similarly, but give a slightly different peak. On real data these differences are more pronounced as will be shown in Section 3.2. Additionally we observe that a Fisher correction for clusters of small size (i.e., less than 8) is necessary to avoid bias in the range of high k, where small clusters would be overweighted (see also Figure 4). The significance of the M-score with respect to permuted data is clear by inspecting profile of the averaged score over 100 permutations, denoted as $Ave(Perm)$ in 1. To illustrate the M-score comprehensively we visualize various matrices A in Figure 2 to show how motif enrichment changes over $k = 2..10$ on synthetic data (binomial method with Fisher correction). We see that, except for the Rap motif, all motifs are detected, though with varying significance (expressed as $-\log(p)$). The Rap motif, which occurs only in two genes in this synthetic setup, is detected only for $k = 4, .., 7$. The overall hidden pattern of motifs is found back for $k = 6$ and $k = 7$. Although displaying the same overall motif pattern, $k = 7$ yields a lower M-score as it contains a cluster void of significant motifs. For $k = 5$ we observe a stronger significance of the Rap motif than for $k = 6$, but a weaker correspondence to the original pattern. This illustrates that our method is of an exploratory nature: it is meant to point to a region of interest rather than to a single 'true' value. Finally, Figure 3 shows the behavior of the M-score on synthetic data for various σ. We see that up to $\sigma = 0.4$, the region around k_{real} is identified. Microarray data is found to typically exhibit a fluctuation of $\sigma = 0.25$ [12].

3.2 Yeast Cell-Cycle Data

To evaluate the performance of the M-score on real data, we use the yeast expression data from Spellman et al.[19]. For computational reasons, we select from the 3000 genes with highest variance, those 1745 genes that had literature annotations in the SGD database[20] as of January 2003. To check whether the choice of this subsample would bias our findings we calculate the linear correlation between the a priori chance to find a motif in this set of genes versus the a priori chance to find a motif over the entire genome. We obtain a correlation of 0.998, ensuring that this gene selection procedure did not drastically change the motif composition in the gene set with respect to analyses in Cho et al.[21] and Tavazoie et al.[22]. Hence, differences in M-scores in the following analyses should not be attributed to the gene selection procedure.

3.2.1 Optimal k in terms of the M-score

To demonstrate the ability of the M-score to evaluate various cluster batches, we cluster the data hierarchically using the cosine distance metric and Ward's method, and take all cutoffs in the range $k = 2, .., 100$. The data was preprocessed as in [23]. Figure 4 plots the M-score over k. We evaluate the score biologically by visualizing the cluster-by-motif matrix A in the cases $k = 5, 13, 30$ in Figure 5.

The clustering with $k = 13$ yields the highest number of significant motifs and generates the most consistent

biological results. It is obvious that 5 clusters (k=5) is not sufficient for clustering the data in a biologically meaningful way. All genes with cell-cycle specific profiles are grouped in the same cluster (cluster 3) regardless of their phase specificity. This explains why in this cluster both the ECB (M/G1 phase specific) and the MCB motif (G1/S phase specific) are enriched. Cluster 2 is characterized by an enrichment of the motifs M1a, M3b, Rap1 and TCS as well as a significant change in expression in the G1 phase. This cluster contains genes that encode ribosomal proteins, characterized by the Rap1 and M1a motif [22] and other proteins involved in translation and RNA metabolism, found with genes carrying the M3b motif [22]. These genes are tightly coregulated since this cluster is also obtained with $k = 13$ (cluster 13) en $k = 30$ (cluster 30), each time containing approximately the same genes. The other three clusters obtained with $k = 5$, are too diverse to yield any enrichment of motifs. These findings were confirmed by visual inspection of the expression profiles of the clusters (results not shown).

The clustering with optimal $k = 13$ leads to five periodic clusters, each specific for a single phase of the cell cycle. In cluster 1 we find an enrichment of M14a and M14b motifs in genes whose expression peaks during the S and the S/G2 transition phase , which corresponds to the results of [22]. Cluster 6 shows an enrichment of the G1/S specific MCB motif [22][19] and the M14a motif, which was thought to be S/G2 specific [22]. The expression profiles of these genes clearly show their specificity for the S phase (Figure 6). This phase-specificity is shared by cluster 5 (Figure 7), which shows an enrichment for binding motifs for the Fkh1, Fkh2 and Ndd1 transcription factors, which are known to cooperate during the S/G2 transition [24]. Cluster 2 is characterized by an enrichment in ECB, Pho4, Cbf1p and Met31-32p motifs. The latter two factors are known to cooperatively regulate gene expression according to the demand for the products of the amino acid biosynthesis pathway, while Pho4 is implied in the biosynthesis of the amino acid histidine [25]. The expression of the genes in this cluster peaks from the S/G2 transition to the G2/M transition (Figure 8), which is consistent with the reported phase-specificity of the Cbf1p and Met31-32p motifs [22]. However, up to now the cell-cycle specific ECB motif was thought to be specific for the M phase [22] or M/G1 boundary [19], and not to cooccur with the Met31-32p and Cbf1p motifs. Finally, the STRE motif is characteristic for cluster 10 and this cluster was also found by [22]. Additionally, five non-periodic clusters were found that showed a significant enrichment of one or several motifs. On the other hand, when the data are clustered into a large number of groups ($k = 30$), the cluster with the combined M14a and M14b motifs disappears and Fkh2 is no longer found in combination with its known interaction partners Fkh1 and Ndd1 [24]. It should be stressed that we only clustered a subset of the genes that were analyzed by Tavazoie et al., therefore it is to be expected that our motif results are different. However, taking into account the fact that we clustered only half of the genes, our results seem very promising considering they only found 3 periodic and 5 relevant non-periodic clusters.

Finally, as already observed in Section 3.1, the overall behavior of the Pearson-based score does not differ qualitatively from using binomial p-values (corrected for small clusters and multiple testing in both cases), apart from the fact that Pearson is less conservative (lower p-values). The Bonferroni correction accommodates for the overestimation of motif significance and thereby improves average-to-peak ratio of the scoring curve. After permuting for each k the cluster assignments 100 times we see in that the permuted data occasionally results in significant motifs (i.e., yielding M-scores > 10), but we find the complete M-score profile over k significant at level p=0.05 (see confidence band in Figure 4).

3.2.2 M-score to compare cluster methods

We illustrate how the proposed heuristic behaves for some other standard clustering methods. Figure 9 shows compares complete linkage and ward's method in hierarchical clustering (cosine distance) with k-means. Single linkage and average linkage are omitted since they did not produce significant clusters in terms of motifs. As k-means is not deterministic, an average profile over 30 runs with different initializations is plotted. We see that k-means reaches optima of the M-score in the range $k = 5, .., 15$, which is in partial correspondence with the results from Section 3.2.1. Nevertheless k-means scores overall better, and stops yielding interesting clusters above $k = 70$. Additionally, a peak is reached for $k = 5$ which is rather unexpected.

To interpret this we visualize in Figure 10 the motif enrichment for k-means in two cases. Comparing the hierarchical (ward) structure for $k = 5$ (see Figure 5 with the corresponding k-means solution for $k = 5$, we clearly observe that k-means produces a richer structure. Juxtaposing ward's solution for $k = 12$ en k-means for $k = 17$ (i.e., the best parameterizations in terms of the M-score for both algorithms), we see that an increased significance associated with stronger groupings of M14a-, M3a-, SCB- and ECB-related genes respectively, build up the difference in M-score. As the score balances between a moderate presence of biologically relevant motifs

on the one hand, and a moderate number of clusters on the other, occasionally biases arise. We find an example in the case of k-means with $k = 5$: although the result might have been reasonable, further exploration of slightly higher k was necessary in this case. Finally, we observe that complete linkage clustering does not appear to an interesting choice in this setup.

4 Conclusion

In this work we presented a motif-based score (M-score) to quickly estimate the validity of multiple clustering runs in terms of significant sequence patterns or motifs. We tested it on synthetic and real data of the yeast cell-cycle. Our purpose in this paper was to treat, verify and demonstrate the score, rather than to exhaustively explore its behavior on various data sets or cluster methodologies. We have shown that starting from binomial p-values is a reasonable approach provided a Fisher correction for small cluster sizes and an overall Bonferroni correction for multiple testing. Our background aim was to economize on biological validations when the parameter-space is prohibitively large. Therefore, we do not consider the M-score an absolute quantification of biological relevance and we illustrated possible biases in an example of k=5. We tested the M-score on two mainstream cluster algorithms using the number of clusters k as variable parameter. In principle, however, also other cluster algorithms with parameterizations different from the *number of clusters*, k can be explored. For that matter we have also experimented with the parameter S in Adaptive Quality-based Clustering [23], which yielded comparable results that did not add extra arguments to this work (results not shown). In another context, we used the score to check how appending or integrating text-based information with microarray data can reveal gene groupings with both functional and motif enrichments that are not detectable by expression data alone[26]. In our setup we demonstrated the score using sets of known motifs, although this is not part of the stated assumptions and the method is also applicable for sets of newly discovered conserved sequences [11]. A general limitation of the method is that is attempts to quantify solutions in terms of *direct* regulation and makes no claim to score genetic networks. Finally, given that the evaluation of clustering solutions in 'motif space' provides a useful alternative to other validation procedures as described in Section 1, we see its relation to such other scores a useful topic of further investigation.

As a conclusion we state that although the biological assumptions that underpin the heuristic constitute a simplification of the biological reality and advanced algorithms that integrate sequence information more deeply already exist, the proposed M-score proves a useful tool to explore the effect of multiple clustering parameterizations and methods in terms of detecting regulatory patterns over an entire data set.

5 Acknowledgements

PG is a research assistant of the KULeuven. JM is a post-doctoral researcher of the KULeuven. YM is a post-doctoral researcher of the FWO and an assistant professor at the K.U.Leuven. BDM is a full professor at the Katholieke Universiteit Leuven, Belgium. This research is supported by : *Research Council KUL*: GOA-Mefisto 666, IDO (IOTA Oncology), several PhDpostdoc and fellow grants; *Flemish Government*: FWO: PhDpostdoc grants, projects G.0115.01 (microarraysoncology), G.0240.99 (multilinear algebra), G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), research communities (ICCoS, ANMMM); IWT: PhD Grants, GBOU-McKnow (Knowledge management algorithms), GBOU-SQUAD (quorum sensing); Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-22 (2002-2006)); EU: CAGE; ERNSI; Contract Research/agreements: Data4s, Electrabel, Elia, LMS, IPCOS, VIB; We wish to thank Stein Aerts for useful feedback on the manuscript.

6 Authors' contributions

PG developed the M-score and carried out the experiments. JM did the biological evaluations. YM and BDM provided overall supervision, feedback and funding of the research.

References

[1] Yeung K, Haynor D, Ruzzo W: **Validating clustering for gene expression data.** *Bioinformatics* 2001, **17**:309–318.

[2] Yeung K, Ruzzo W: **Principal component analysis for clustering gene expression data**. *Bioinformatics* 2001, **17**:763–774.

[3] Kaufman L, Rousseeuw P: *Finding groups in data*. Wiley-Interscience, 1990.

[4] Pollard K, van der Laan M: **A method to identify significant clusters in gene expression data**. In *To appear in Proc of Systemics, Cybernetics and Informatics 2002 (SCI 2002)*, 2002 .

[5] Tibshirani R, Walther G, Hastie T: **Estimating the number of clusters in a dataset via the gap statistics**. *Tech. Rep. 208*, Stanford University, USA, 2000.

[6] Ben-Hur A, Elisseeff A, Guyon I: **A stability based method for discovering structure in clustered data**. In *Proc of the Seventh Ann Pac Symp Biocomp (PSB 2002)*, 2002 6–17.

[7] Zeeberg B, Feng W, Wang G, Wang M, Fojo A, Sunshine M, Narasimhan S, Kane D, Reinhold W, Lababidi S, *et al.*: **GOMINER: A Resource for Biological Interpretation of Genomic and Proteomic Data**. *Genome Biol* 2003, **4(4):R28**.

[8] Segal E, Shapira M, Rev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data**. *Nat Genet* 2003, **34**:166–176.

[9] Coessens B, Thijs G, Aerts S, Marchal K, De Smet F, Engelen K, Glenisson P, Moreau Y, Mathys J, De Moor B: **INCLUSive: A web portal and service registry for microarray and regulatory sequence analysis.** *Nucleic Acids Res* 2003, **31**:3468–3470.

[10] Glenisson P, Coessens B, Van Vooren S, Moreau Y, De Moor B: **Text-based gene profiling with domain-specific views**. In *Proc. of the First International Workshop on Semantic Web and Databases (SWDB 2003), Berlin, Germany*, 2003 15–31.

[11] Vilo J, Brazma A, Jonassen I, Robinson A, Ukkonen E: **Mining for putative regulatory elements in the yeast genome using expression data**. In *Proc of the Eighth International Conference on Intelligent Systems for Molecular Biology*, 2000 384–394.

[12] Bussemaker H, Li H, Siggia E: **Regulatory element detection using correlation with expression**. *Nat Genet* 2001, **27**:167–171.

[13] Conlon E, Liu X, Lieb J, Liu J: **Integrating regulatory motif discovery and genome-wide expression analysis**. *Proc Natl Acad Sci USA* 2003, **100**:3339–3344.

[14] Gasch A, Eisen M: **Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering**. *Genome Biol* 2002, **3**:1–22.

[15] Jobson J: *Applied Multivariate Data Analysis, v.2. Categorical and multivariate methods*. Springer-Verlag, 1992.

[16] Lowry R: **Concepts and Applications of Inferential Statistics** [http://faculty.vassar.edu/lowry/VassarStats.html].

[17] Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel A, Kel-Margoulis O, *et al.*: **TRANSFAC: transcriptional regulation, from patterns to profiles**. *Nucleic Acids Res* 2003, **31**:374–378.

[18] Hughes J, Estep P, Tavazoie S, Church G: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae**. *J Mol Biol* 2000, **296**:1205–1214.

[19] Spellman P, Sherlock G, Zhang M, Iyer V, Anders K, Eisen M, Brown P, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.** *Mol Biol Cell* 1998, **9**:3273–3297.

[20] **Saccharomyces Genome Database**
[http://www.yeastgenome.org].

[21] Cho R, Campbell M, Winzeler E, Steinmetz L, Conway A, Wodicka L, Wolfsberg T, Gabrielian A, Landsman D, Lockhart D, *et al.*: **A genome-wide transcriptional analysis of the mitotic cell cycle.** *Mol Cell* 1998, **2**:65–73.

[22] Tavazoie S, Hughes J, Campbell M, Cho R, Church G: **Systematic determination of genetic network architecture**. *Nat Genet* 1999, **22**:281–285.

[23] De Smet F, Mathys J, Marchal K, Thijs G, De Moor B, Moreau Y: **Adaptive quality-based clustering of gene expression profiles**. *Bioinformatics* 2002, **18**:735–746.

[24] Lee T, Rinaldi N, Robert F, Odom D, Bar-Joseph Z, Gerber G, Hannett N, Harbison C, Thompson C, Simon I, *et al.*: **Transcriptional Regulatory Networks in Saccharomyces cerevisiae.** *Science* 2002, **298**:799–804.

[25] Bhoite L, Allen J, Garcia E, Thomas L, ID IG, WP WV, Whelihan K, Rolfes R, Stillman D: **Mutations in the pho2 (bas2) transcription factor that differentially affect activation with its partner proteins bas1, pho4, and swi5.** *J Biol Chem* 2002, **277**:37612–37618.

[26] Glenisson P, Mathijs J, Moreau Y, Moor BD: **Meta-clustering of gene expression data and literature-extracted information**. *Tech. Rep. 03-162*, ESAT-SCD, K.U.Leuven, Belgium, 2003.

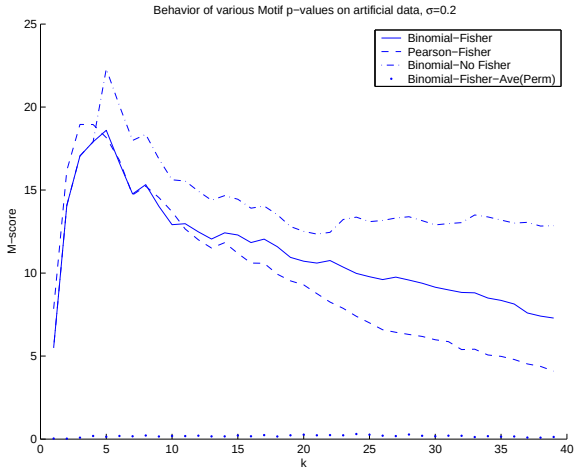

Figure 1: M-score versus k on synthetic data. We show the influence of various p-values on the M-score. The real number of clusters, $k_{real} = 6$, is correctly identified by the binomial-related methods. A Fisher correction for clusters of small size (i.e., less than 8) is necessary to avoid bias in the range of high k. The averaged score over 100 permutations is denoted as $Ave(Perm)$.

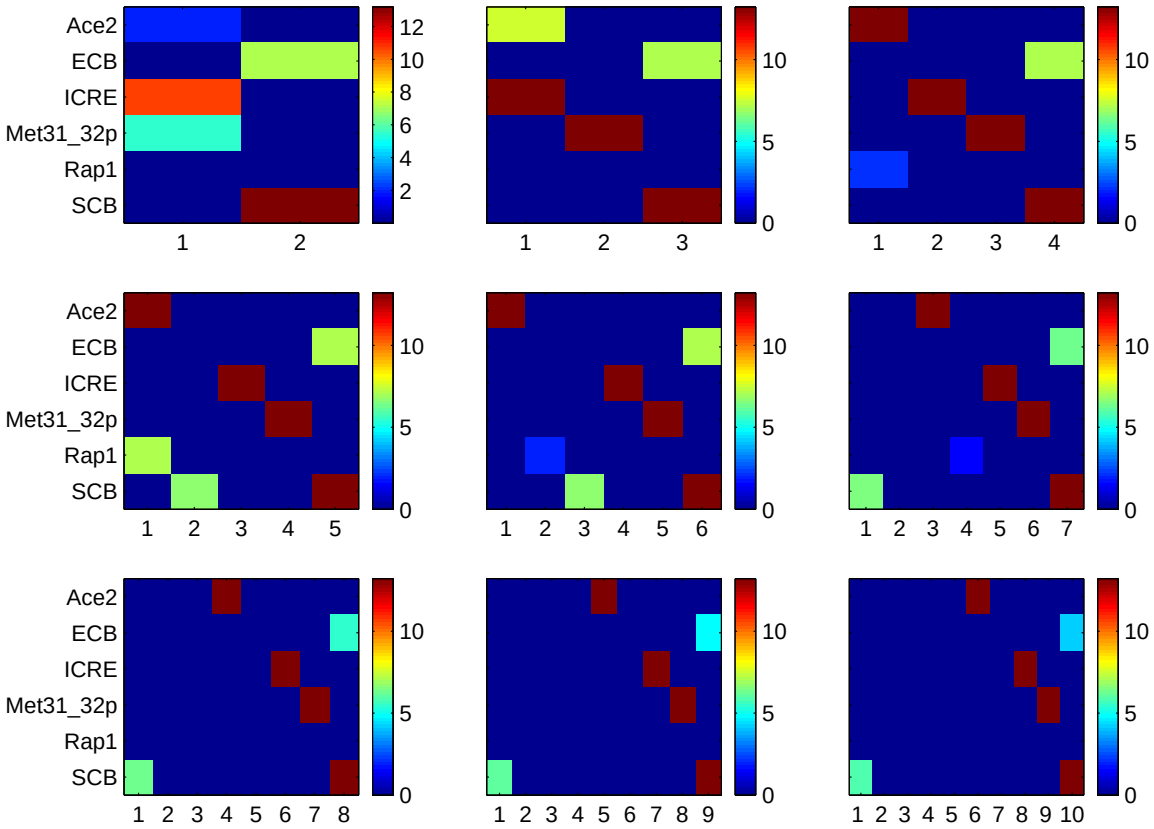

Figure 2: Visualization of motif enrichment matrix A for various k on synthetic data ($\sigma = 0.2$). The colorbar indicates the negative logarithm of p-values which stem from the binomial procedure with a Fisher correction. The Rap motif, which occurs only in two genes in this synthetic setup, is detected only for $k = 4, .., 7$. The overall hidden pattern of motifs is found back for $k = 6$ and $k = 7$.

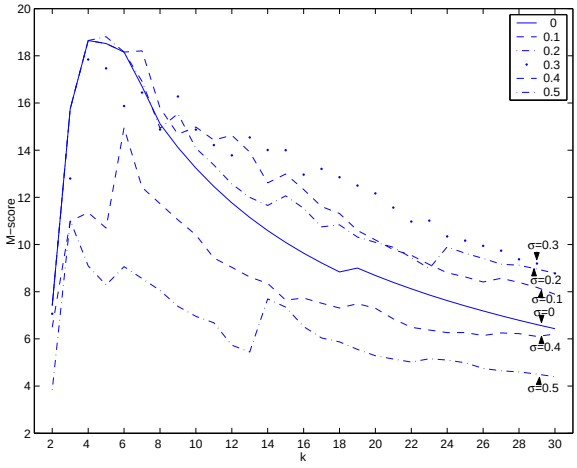

Figure 3: M-score versus k on synthetic data with various σ. We illustrate how robust the M-score is with respect to noise added $\mathcal{N}(0, \sigma)$. Up to $\sigma = 0.4$, the score identifies the region around k_{real} as the one of interest.

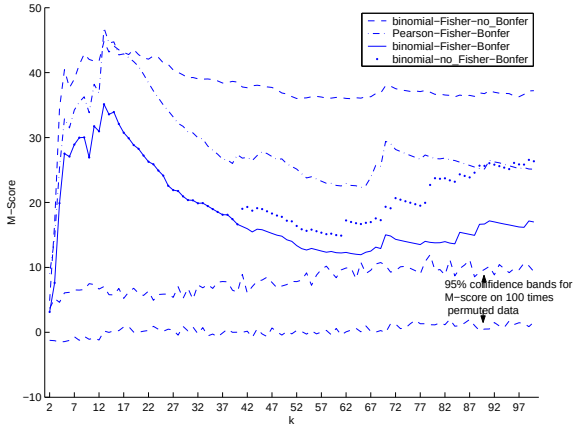

Figure 4: M-score versus k on yeast expression data. The Fisher correction adjusts overestimation of p-values in small clusters that are typical for high k. The Bonferroni correction accommodates for the overestimation of motif significance and thereby improves average-to-peak ratio of the scoring curve. The M-score is significant with respect to permuted data.

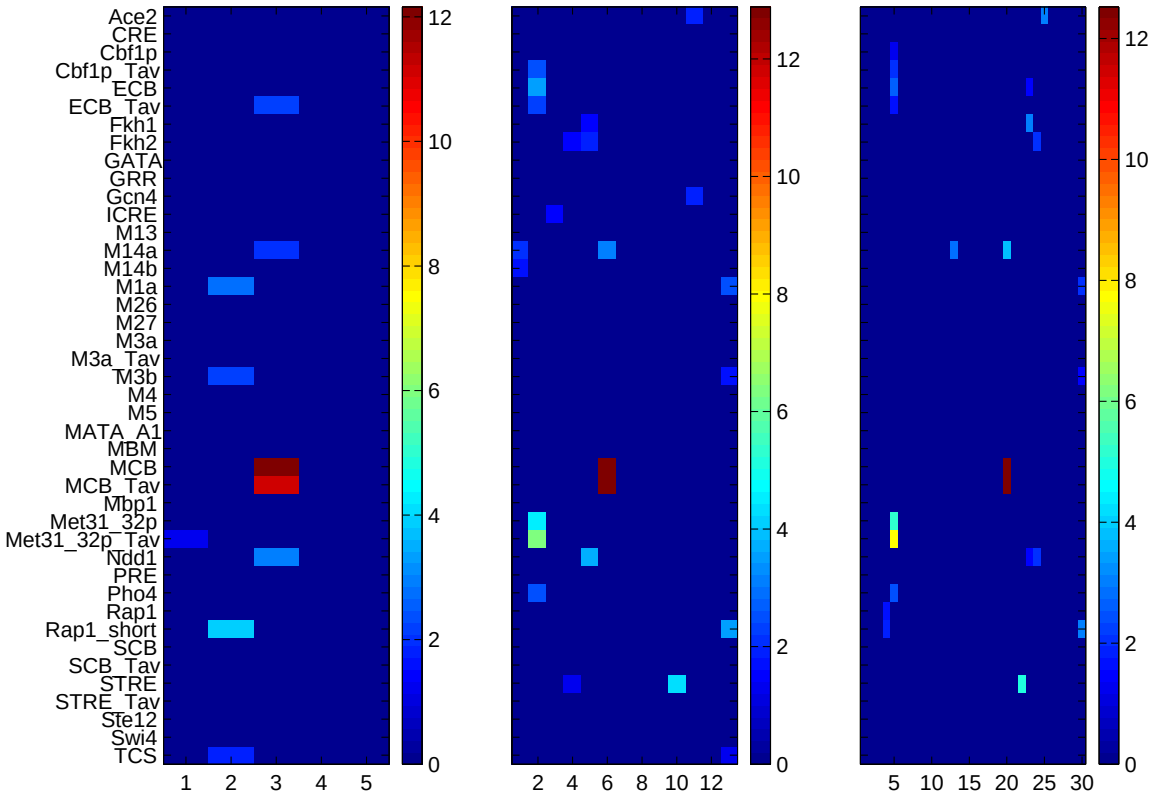

Figure 5: Visualization of motif enrichment matrices A for ward's hierarchical clustering on cell-cycle data with various k.

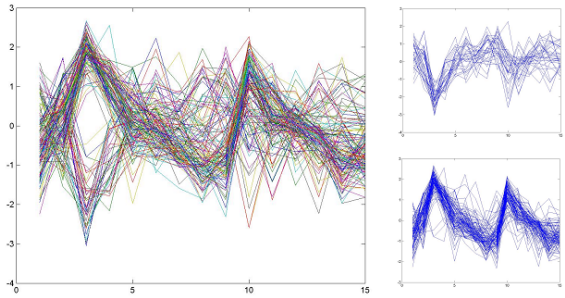

Figure 6: Expression profiles of cluster 6, enriched with M14a and MCB motifs.We append a discoupled view of the up- and down-regulated profiles for clarity.

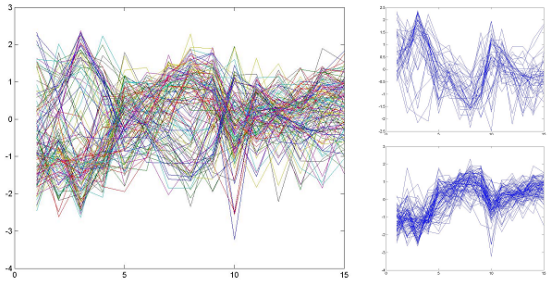

Figure 7: Expression profiles of cluster 5, enriched with Fkh1, Fkh2 and Ndd1 motifs.We append a discoupled view of the up- and down-regulated profiles for clarity.

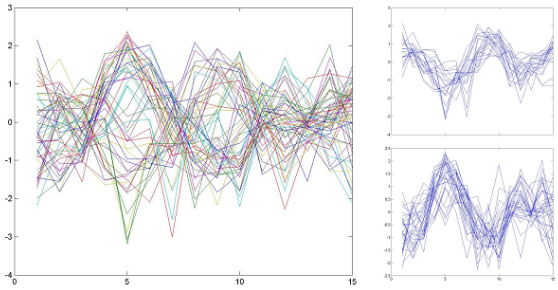

Figure 8: Expression profiles of cluster 2, enriched with ECB, Pho4, Cbf1p and Met31-32p motifs. We append a discoupled view of the up- and down-regulated profiles for clarity.

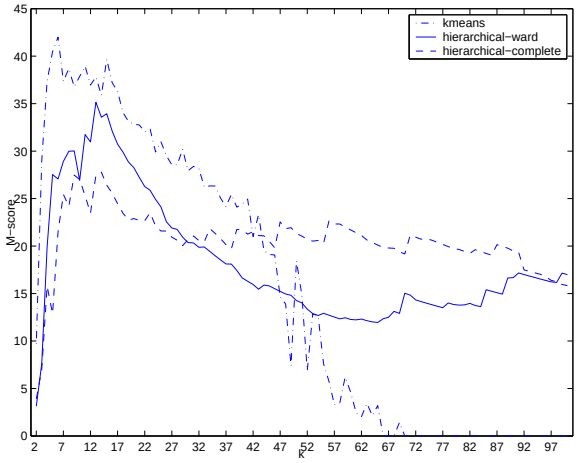

Figure 9: M-score versus k for various basic cluster algorithms. All M-score profiles are generated with binomial p-values, with the correction for small clusters (Fisher method) and multiple testing

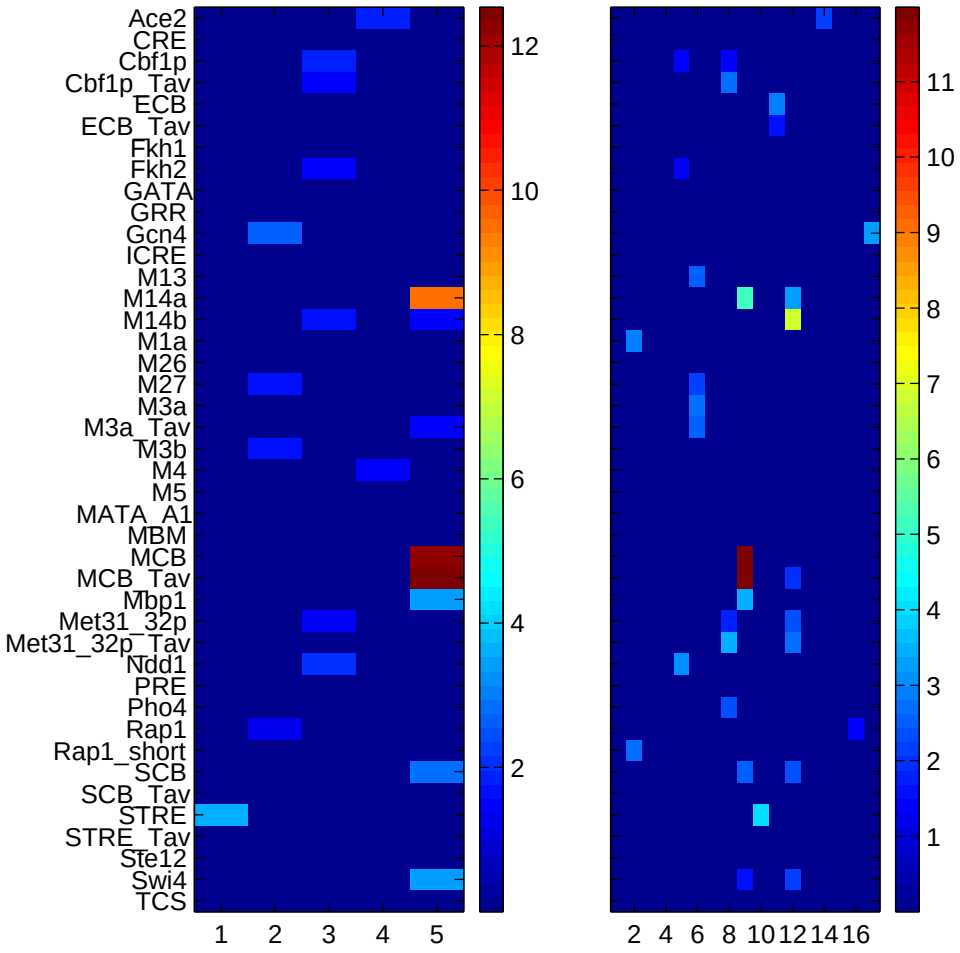

Figure 10: Visualization of motif enrichment matrices A for k-means on cell-cycle data with $k = 5$ and $k = 17$

Table 1: Contigency Table C. Elements of this table in the text are denoted by c_{ij} $(i, j = 1, 2)$

	Motif	¬Motif
Cluster	r	$n - r$
¬Cluster	$m - r$	$N - n - m + r$

Table 2: Partitioning of 358 genes bearing the six selected motifs. Ninety-four genes carrying the SCB motif are merged with the ECB genes.

Group	Motifs(nr_genes)
1	Ace2(19)
2	ECB(15)+SCB(94)
3	ICRE(75)
4	Met31_32p(52)
5	Rap1(2)
6	SCB(101)

Table 3: Additional Table: Motifs considered in cell-cycle experiment. The consensus sequence for each motif was extracted from the TRANSFAC Database

Motif	sequence
Ace2	GTGTGTGTGTGTGTG
Cbf1p_Tav	TCACGTG
Cbf1p	RTCACRTG
CRE	TKACGTCA
ECB_Tav	TTWCCNNNWNNGGAAA
ECB	TTWCCCNNWNAGGAAA
Fkh1	AAWNRTAAAYARGN
Fkh2	GGNAAWANGTAAACAA
GATA	GATAA
Gcn4	TGACTCA
GRR	CTTCTT
ICRE	WYTTCAYRNTGS
M13	WWWTTTTTCA
M14a	TTTSGCKTY
M14b	TGTTTRKT
M1a	TTCYCWYYG
M26	GTKTNNGTGNG
M27	AWATTTTTC
M3a_Tav	MGATGAG
M3a	CGATGAG
M3b	TGAAAAA
M4	GAARKYTTT
M5	YWKNTCCYNTTT
MATA_A1	TGATGTANNT
MBM	TATTTGCTC
Mbp1	GNRRNANACGCGTNR
MCB_Tav	ACGCGKNR
MCB	ACGCGTNA
Met31_32p_Tav	AAANTGTGG
Met31_32p	AAACTGTGG
Ndd1	GGNAAWWNGTMAACA
Pho4	CACGTG
PRE	TGAAACA
Rap1_short	ACCCAT
Rap1	ACACCCATACATTT
SCB_Tav	MAMGAAA
SCB	CACGAAAA
Ste12	TGAAACA
STRE_Tav	AGGRG
STRE	AGGGG
Swi4	CACGAAA
TCS	CATTCT