BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 9 2006, pages 1122–1129
doi:10.1093/bioinformatics/btl060

Gene expression

A systematic comparison and evaluation of biclustering methods for gene expression data

Amela Prelić[1], Stefan Bleuler[1,*], Philip Zimmermann[2], Anja Wille[3,4], Peter Bühlmann[4], Wilhelm Gruissem[2], Lars Hennig[2], Lothar Thiele[1] and Eckart Zitzler[1]

[1]Computer Engineering and Networks Laboratory, [2]Institute for Plant Sciences and Functional Genomics Center Zurich, [3]Colab and [4]Seminar for Statistics, ETH Zurich, 8092 Zurich, Switzerland

Received on July 27, 2005; revised on January 4, 2006; accepted on February 15, 2006
Advance Access publication February 24, 2006
Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: In recent years, there have been various efforts to overcome the limitations of standard clustering approaches for the analysis of gene expression data by grouping genes and samples simultaneously. The underlying concept, which is often referred to as biclustering, allows to identify sets of genes sharing compatible expression patterns across subsets of samples, and its usefulness has been demonstrated for different organisms and datasets. Several biclustering methods have been proposed in the literature; however, it is not clear how the different techniques compare with each other with respect to the biological relevance of the clusters as well as with other characteristics such as robustness and sensitivity to noise. Accordingly, no guidelines concerning the choice of the biclustering method are currently available.

Results: First, this paper provides a methodology for comparing and validating biclustering methods that includes a simple binary reference model. Although this model captures the essential features of most biclustering approaches, it is still simple enough to exactly determine all optimal groupings; to this end, we propose a fast divide-and-conquer algorithm (Bimax). Second, we evaluate the performance of five salient biclustering algorithms together with the reference model and a hierarchical clustering method on various synthetic and real datasets for *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. The comparison reveals that (1) biclustering in general has advantages over a conventional hierarchical clustering approach, (2) there are considerable performance differences between the tested methods and (3) already the simple reference model delivers relevant patterns within all considered settings.

Availability: The datasets used, the outcomes of the biclustering algorithms and the Bimax implementation for the reference model are available at http://www.tik.ee.ethz.ch/sop/bimax

Contact: bleuler@tik.ee.ethz.ch

Supplementary information: Supplementary data are available at http://www.tik.ee.ethz.ch/sop/bimax

INTRODUCTION

In recent years, several biclustering methods have been suggested to identify local patterns in gene expression data. In contrast to classical clustering techniques such as hierarchical clustering (Sokal and Michener, 1958) and *k*-means clustering (Hartigan and Wong, 1979), biclustering does not require genes in the same cluster to behave similarly over all experimental conditions. Instead, a bicluster is defined as a subset of genes that exhibit compatible expression patterns over a subset of conditions. This modified clustering concept can be useful to uncover processes that are active only over some but not all samples as has been demonstrated in several studies (Cheng and Church, 2000; Ihmels *et al*., 2002; Ben-Dor *et al*., 2002; Tanay *et al*., 2002; Murali and Kasif, 2003), see Madeira and Oliveira (2004) for a survey.

Comparing clustering methods in general is difficult as the formalization in terms of an optimization problem strongly depends on the scenario under consideration and accordingly varies for different approaches. In the end, biological merit is the main criterion for validation, though it can be intricate to quantify this objective. In the literature, there are several comparative studies on traditional clustering techniques (Yeung *et al*., 2001; Azuaje, 2002; Datta and Datta, 2003); however, for biclustering no such extensive empirical comparisons exist as pointed out by (Madeira and Oliveira (2004). Although first steps in this directions have been taken (Tanay *et al*., 2002; Yang *et al*., 2003; Ihmels *et al*., 2004), the corresponding studies focus on validating a new algorithm with regard to one or two existing biclustering methods and usually consider a specific objective function.

The main goal of this paper is to provide a systematic comparison and evaluation of prominent biclustering methods in the light of gene classification. In particular, we focus on the following questions: (1) What comparison/validation methodology is adequate for the biclustering context, (2) how meaningful are the biclusters selected by existing methods and (3) how do different methods compare with each other, i.e. do some techniques have advantages over others or are there common properties that all approaches share?

In order to answer these questions, we have selected a number of salient biclustering methods, implemented them and tested them on both synthetic and real gene expression datasets. An *in silico* scenario has been chosen to (1) investigate the capability of the algorithms to recover implanted transcription modules (Ihmels *et al*., 2002), i.e. sets of co-regulated genes together with their regulating conditions and (2) study the influence of regulatory complexity and noise on the performance of the algorithms. To assess the biological relevance of biclusters on gene expression data for *Saccharomyces cerevisiae* and *Arabidopsis thaliana*,

*To whom correspondence should be addressed.

multiple quantitative measures are introduced that relate the biclustering outcomes to annotations by Gene Ontology Consortium (2000) metabolic pathway maps and protein-interaction data.

Moreover, we propose a simple biclustering model, which retains common features of most biclustering methods, in combination with a fast and exact algorithm (Bimax)—in contrast, existing biclustering algorithms usually do not guarantee to find global optima. Although restricted from a biological point of view, this model allows to study the validity of the biclustering idea independent of the interfering effects because of approximate algorithms. As such, Bimax has been considered as a reference method in our study. As will be shown in the remainder of this paper, even such a simple approach delivers biologically relevant results and compares well with more sophisticated biclustering methods.

RELATED WORK

There exist several studies that address the issue of comparing and validating one-dimensional clustering methods (Kerr and Churchill, 2001; Yeung *et al*., 2001; Azuaje, 2002; Datta and Datta, 2003; Gat-Viks *et al*., 2003; Handl *et al*., 2005). All of them make use of different quantitative measures or validity indices, which can be divided into three categories (Halkidi *et al*., 2001): internal, external and relative indices. Internal indices solely rely on the input data as, e.g. the measures of homogeneity and separation (Gat-Viks *et al*., 2003). In contrast, external criteria are based on additional data in order to validate the obtained results. In the context of gene expression data, these would correspond to prior biological knowledge of the systems being studied; alternatively, a validation can be done by referring to other types of genomic data representing similar aspects of the regulation mechanisms being investigated. The third category of relative indices measures the influence of the input parameter settings on the clustering outcome. As discussed in Handl *et al*. (2005), external indices are preferable in order to assess the performance of an algorithm on a given dataset, while internal indices can be used to investigate why a particular method does not perform well.

In the context of biclustering, mainly external validation has been used. Biological analyses and interpretations by human experts are most common for the evaluation of a single, newly proposed biclustering algorithm (Cheng and Church, 2000; Getz *et al*., 2000; Ben-Dor *et al*., 2002; Murali and Kasif, 2003; Bergmann *et al*., 2003; Getz *et al*., 2003; Ihmels *et al*., 2004); they are usually descriptive and qualitative only, and therefore are not suitable for comparing multiple methods. In terms of quantitative measures, many papers rely on known classifications and categorizations given by tumor types (Tanay *et al*., 2002; Kluger *et al*., 2003; Murali and Kasif, 2003), GO annotations (Tanay *et al*., 2002, Tany *et al*., 2004), metabolic pathways (Ihmels *et al*., 2002) or promoter motifs (Ihmels *et al*., 2004), which are closely related to the specific datasets under consideration. Some authors also investigate *in silico* datasets with implanted biclusters where the optimal outcome is known beforehand (Ihmels *et al*., 2002; Ben-Dor *et al*., 2002; Bergmann *et al*., 2003; Yang *et al*., 2002).

Most biclustering papers are concerned with the introduction and validation of a new approach, while only a few contain quantitative comparisons with existing methods. Cheng and Church (2000) and Kluger *et al*. (2003), validate the biclustering results in comparison with hierarchical clustering and singular value decomposition respectively. Tanay *et al*. (2002) and Yang *et al*. (2002, 2003), provide a comparison with the algorithm by Cheng and Church, (2000) regarding synthetic data respectively the problem formulation introduced in Cheng and Church (2000). In Ihmels *et al*. (2004), two biclustering techniques (Cheng and Church, 2000; Getz *et al*., 2000) as well as five classical clustering methods are tested with respect to the problem formulation used by the iterative signature algorithm proposed in Ihmels *et al*. (2002). In most of the studies, the comparison has been carried out with regard to the gene dimension.

BICLUSTERING METHODS

Selected algorithms

Five prominent biclustering methods have been chosen for this comparative study according to three criteria: (1) to what extent the methods have been used or referenced in the community, (2) whether their algorithmic strategies are similar and therefore better comparable and (3) whether an implementation was available or could be easily reconstructed based on the original publications. The selected algorithms, which all are based on greedy search strategies, are Cheng and Church's algorithm CC (Cheng and Church, 2000); Samba (Tanay *et al*., 2002); Order Preserving Submatrix Algorithm, OPSM (Ben-Dor *et al*., 2002); Iterative Signature Algorithm, ISA (Ihmels *et al*., 2002, 2004); *xMotif* (Murali and Kasif, 2003). A brief description of the corresponding approaches can be found in the Supplementary Material.

Reference method (Bimax)

The above methods use different models which are all too complex to be solved exactly; most of the corresponding optimization problems have shown to be NP-hard. Therefore, advantages of one method over another can be due to a more appropriate optimization criterion or a better algorithm.

To decouple these two aspects, we propose a reference method, namely Bimax, that uses a simple data model reflecting the fundamental idea of biclustering, while allowing to determine all optimal biclusters in reasonable time. This method has the benefit of providing a basis to investigate (1) the usefulness of the biclustering concept in general, independently of interfering effects caused by approximate algorithms, and (2) the effectiveness of more complex scoring schemes and biclustering methods in comparison to a plain approach. Note that the underlying binary data model, which is described below, is only used by Bimax and does not represent the platform on the basis of which the different algorithms are compared. All methods under consideration are employed using their specific data models.

Model The model assumes two possible expression levels per gene: no change and change with respect to a control experiment.[1] Accordingly, a set of m microarray experiments for n genes can be represented by a binary matrix $E^{n \times m}$, where a cell e_{ij} is 1 whenever gene i responds in the condition j and otherwise it is 0. A bicluster (G, C) corresponds to a subset of genes $G \subseteq \{1, \ldots, n\}$ that jointly

[1]To this end, a preprocessing step normalizes log expression values and then transforms matrix cells into discrete values, e.g. by using a 2-fold change cutoff.

respond across a subset of samples $C \subseteq \{1, \ldots, m\}$. In other words, the pair (G, C) defines a submatrix of E for which all elements equal 1. Note that, by definition, every cell e_{ij} having value 1 represents a bicluster by itself. However, such a pattern is not interesting *per se*; instead, we would like to find all biclusters that are inclusion-maximal, i.e. that are not entirely contained in any other bicluster.

DEFINITION 1. *The pair* $(G, C) \in 2^{\{1,\ldots,n\}} \times 2^{\{1,\ldots,m\}}$ *is called an inclusion-maximal bicluster if and only if* (1) $\forall\, i \in G, j \in C : e_{ij} = 1$ *and* (2) $\nexists (G', C') \in 2^{\{1,\ldots,n\}} \times 2^{\{1,\ldots,m\}}$ *with* (a) $\forall i' \in G', j' \in C'$: $e_{ij'} = 1$ *and* (b) $G \subseteq G' \wedge C \subseteq C' \wedge (G', C') \neq (G, C)$.

This model is similar to the one presented in Tanay *et al*. (2002) where a bicluster can also contain 0-cells.

Algorithm Since the size of the search space is exponential in n and m, an enumerative approach is infeasible in order to determine the set of inclusion-maximal biclusters. Alexe *et al*. (2002) proposed an algorithm in a graph-theoretic framework that can be employed in this context, if the matrix E is regarded as an adjacency matrix of a graph. By exploiting the fact that the graph induced by E is bipartite, their incremental algorithm can be tailored to this application which reduces the running-time complexity from $\Theta(n^2 m^2 \beta)$ to $\Theta(nm\beta \log \beta)$, where β is the number of all inclusion-maximal biclusters in $E^{n \times m}$ (see Supplementary Material). However, the memory requirements of this algorithm are of order $\Omega(nm\beta)$ which causes practical problems, especially for larger matrices.

In this paper, though, we propose and use a fast divide-and-conquer approach, the binary inclusion-maximal biclustering algorithm (Bimax) that requires much less memory resources ($O(nm \min\{n, m\})$), while providing a worst-case running-time complexity that for matrices containing disjoint biclusters only is of order $O(nm\beta)$ and for arbitrary matrices is of order $O(nm\beta \min\{n, m\})$. The complete algorithm and the proof of the general upper bound for the running-time complexity are given in the Supplementary Material. Bimax tries to identify areas of E that contain only 0s and therefore can be excluded from further inspection. This strategy is especially beneficial for our purposes as E is, depending on the cutoff threshold, sparse; in all datasets used in this study, the proportion of 1-cells over 0-cells never exceeded 6% when considering a 2-fold change cutoff.

More specifically, the idea behind the Bimax algorithm, which is illustrated in Figure 1, is to partition E into three submatrices, one of which contains only 0-cells and therefore can be disregarded in the following. The algorithm is then recursively applied to the remaining two submatrices U and V; the recursion ends if the current matrix represents a bicluster, i.e. contains only 1s. If U and V do not share any rows and columns of E, i.e. G_W is empty, the two matrices can be processed independently from each other. However, if U and V have a set G_W of rows in common as shown in Figure 1, special care is necessary to only generate those biclusters in V that share at least one common column with C_V.

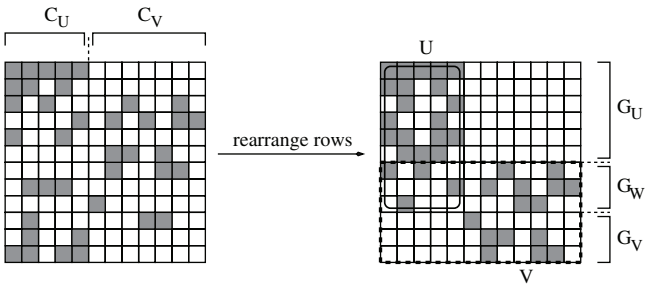

Fig. 1. Illustration of the Bimax algorithm. To divide the input matrix into two smaller, possibly overlapping submatrices U and V, first the set of columns is divided into two subsets C_U and C_V, here by taking the first row as a template. Afterwards, the rows of E are resorted: first come all genes that respond only to conditions given by C_U, then those genes that respond to conditions in C_U and in C_V and finally the genes that respond to conditions in C_V only. The corresponding sets of genes G_U, G_W and G_V then define in combination with C_U and C_V the resulting submatrices U and V which are decomposed recursively.

COMPARISON METHODOLOGY

In general, a fair comparison of clustering and biclustering approaches is inherently a difficult task because every method uses a different problem formulation and algorithm which may work well in certain scenarios and fail in others. Here, the main goal is to define a common setting that reflects the general basis of the majority of the biclustering studies available and in particular of those techniques considered in this paper.

First, the comparison focuses on the identification of (locally) co-expressed genes as in Cheng and Church (2000), Tanay *et al*. (2002), Ben-Dor *et al*. (2002), Ihmels *et al*. (2002, 2004) and Tanay *et al*. (2004). Classification of samples or inference of regulatory mechanisms may be other tasks for which biclustering can be used; however, considering mainly the gene dimension has the advantage of various available annotations—in contrast to the condition dimension—and of the possibility to compare the results with classical clustering techniques.

Second, external indices are used to assess the methods under consideration as in most biclustering papers. The reasons are: (1) it is not clear how to extend notions such as homogeneity and separation (Gat-Viks *et al*., 2003) to the biclustering context (to our best knowledge, no general internal indices have been suggested so far for biclustering) and (2) there are some issues with internal measures, owing which Gat-Viks *et al*. (2003) and Handl *et al*. (2005) recommend external indices for evaluating the performance of (bi)clustering methods. We consider both synthetic and real datasets for the performance assessment. Only the latter allow reliable statements about the biological usefulness of a specific approach, and further biological data, namely GO annotations, as in Tanay *et al*. (2002), Tany *et al*. (2004), metabolic pathways maps, similarly to Ihmels *et al*. (2002) and protein–protein interactions, are used here. In contrast, the former datasets inherently reflect only certain aspects of biological reality, but they have the advantage that the optimal solutions are known beforehand and that the complexity can be controlled and arbitrarily scaled to different levels.

Finally, various biclustering concepts and structures can be considered when using *in silico* data; Madeira and Oliveira (2004) propose several categories on the basis of which they classify existing biclustering approaches. Here, we investigate two types of bicluster concepts: biclusters with constant expression values and biclusters following an additive model where the expression values are varying over the conditions. The former type can be used to test methods designed to identify—according to the terminology in Madeira and Oliveira (2004)—biclusters with constant and coherent values, while the latter type, where the

expression values show the same trend for all genes included, serves as a basis to validate algorithms tailored to biclusters with coherent values and coherent evolutions. Concerning the biclustering structure, we consider two scenarios: (1) multiple biclusters without any overlap in any dimension and (2) multiple biclusters with overlap.

Validation using synthetic data

The artificial model used to generate synthetic gene expression data is similar to an approach proposed by Ihmels *et al*. (2002). In this setting, biclusters represent transcription modules; these modules are defined by (1) a set G of genes regulated by a set of common transcription factors and (2) a set C of conditions in which these transcription factors are active. In the first considered scenario, 10 non-overlapping transcription modules, each extending over 10 genes and 5 conditions, emerge. Each gene is regulated by exactly one transcription factor and in each condition only one transcription factor is active. The corresponding datasets contain 10 implanted biclusters and have been used to study the effects of noise on the performance of the biclustering methods. For the second scenario, the regulatory complexity has been systematically varied: here, each gene can be regulated by d transcription factors and in each condition up to d transcription factors can be active. As a consequence, the original 10 biclusters overlap where d is an indicator for the overlap degree; overall, 9 different levels have been considered with $d = 0, 1, \ldots, 8$. Moreover, we have investigated for each scenario two types of biclusters: (1) constant biclusters and (2) additive biclusters (see Supplementary Material).

In order to assess the performance of the selected biclustering approaches, we will use the following gene match score.

DEFINITION 2. *Let M_1, M_2 be two sets of biclusters. The* gene match score *of M_1 with respect to M_2 is* given by the function

$$S^*_G(M_1, M_2) = \frac{1}{|M_1|} \sum_{(G_1, C_1) \in M_1} \max_{(G_2, C_2) \in M_2} \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|}$$

which reflects the average of the maximum match scores for all biclusters in M_1 with respect to the biclusters in M_2.

Now, let M_{opt} denote the set of implanted biclusters and M the output of a biclustering method. The average bicluster relevance is defined as $S^*_G(M, M_{\text{opt}})$ and reflects to what extent the generated biclusters represent true biclusters in the gene dimension. In contrast, the average module recovery, given by $S^*_G(M_{\text{opt}}, M)$, quantifies how well each of the true biclusters is recovered by the biclustering algorithm under consideration. Both scores take the maximum value of 1, if $M_{\text{opt}} = M$. A detailed description of this score can be found in the Supplementary Material.

Validation using prior knowledge

Prior biological knowledge in the form of natural language descriptions of functions and processes that genes are related to has become widely available. One of the largest organized collection of gene annotations is currently provided by Gene Ontology Consortium (2000). Similar to the idea pursued in Tanay *et al*. (2002), we here investigate whether the groups of genes delivered by the different algorithms show significant enrichment with respect to a specific Gene Ontology (GO) annotation. In detail, biclusters are evaluated by computing the hypergeometric functional enrichment score, cf. (Berriz *et al*., 2003), based on Molecular Function and Biological Process annotations; the resulting scores are adjusted for multiple testing by using the Westfall and Young procedure (Westfall and Young, 1993; Berriz *et al*., 2003). This analysis is performed for the model organism *S.cerevisiae*, since the yeast GO annotations are more extensive compared to other organisms. The gene expression dataset used is the one provided by Gasch *et al*., 2000, which contains a collection of 173 different stress conditions and a selection of 2993 genes.

In addition to GO annotations, we consider specific biological networks, namely metabolic and protein–protein interaction networks, that have been derived from other types of data than gene expression data. Although each type of data reveals other aspects of the underlying biological system, one can expect to a certain degree that genes that participate in the same pathway respectively form a protein complex also show similar expression patterns as discussed in Zien *et al*. (2000), Ideker *et al*. (2002), Ihmels *et al*. (2002). The question here is whether the computed biclusters reflect this correspondence.

To this end, we model both pathway information as well as protein interactions in terms of an undirected graph where a node stands for a protein and an edge represents a common reaction in that the two connected proteins participate respectively a measured interaction between the two connected proteins. In order to verify whether a given bicluster (G, C) is plausible with respect to the metabolic respectively protein interaction graph, we consider two scores: (1) the proportion of pairs of genes in G for which there exists no connecting path in the graph and (2) the average path length of pairs of genes in G for which such a path exists. One may expect that both the number of disconnected gene pairs and the average distance between two connected genes is significantly smaller for genes in G than for randomly chosen genes. For both scores, a resampling method is applied where 1000 random gene groups of the same size as G are considered; a Z-test is used to check whether the scores for the bicluster (G, C) are significantly smaller or larger than the average score for the random gene groups.

As to the metabolic level, we use a pathway map that describes the main bio-synthetic pathways at the level of enzymatic reactions for the model organism *A.thaliana* (Wille *et al*., 2004). As this map has been manually assembled at the Institute for Plant Science at ETH Zurich by an extensive literature search, the resulting graph represents a high level of confidence. The gene expression dataset used in this context are publicly available at http://nasc.nott.ac.uk/ and comprise 69 experimental conditions and a selection of 734 genes.

To investigate the correspondence of biclusters and protein–protein interaction networks, again *S.cerevisiae* is considered because the amount of interaction data available is substantially larger than for *A.thaliana*. Here, we combine the aforementioned dataset for yeast (Gasch *et al*., 2000) with protein interactions stored in the DIP database (Salwinski *et al*., 2004), resulting in 11 498 interactions for 3665 genes overall.

Implementation issues

All of the selected methods have been re-implemented according to the specifications in the corresponding papers, except of Samba for which a publicly available software tool, Expander (Sharan *et al*., 2003), has been used. The OPSM algorithm has been slightly

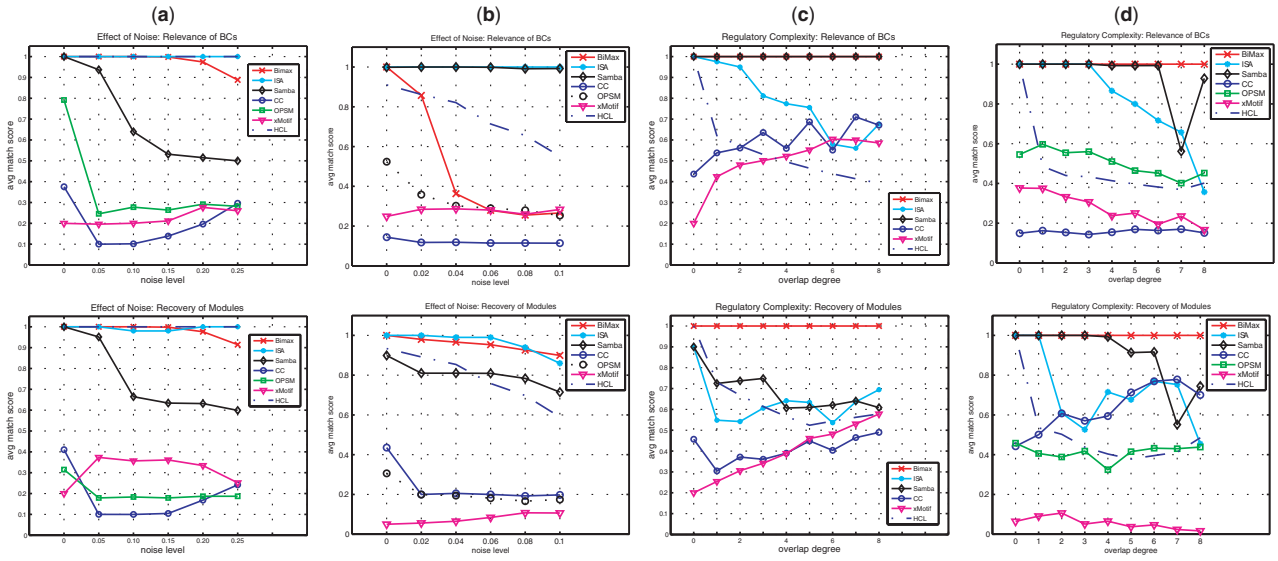

Fig. 2. Results for the artificial scenarios: non-overlapping modules with increasing noise levels for (**a**) constant and (**b**) additive biclusters, overlapping modules with increasing overlap degree and no noise for (**c**) constant and (**d**) additive biclusters. Note that OPSM is excluded in (c), cf. results section.

extended to return not only a single bicluster but also the q largest biclusters among those that achieve the optimal score; q has been set to 100. Furthermore, the standard hierarchical clustering method (HCL) in MATLAB has been included in the comparison, which uses single linkage in combination with Euclidean distance. The parameter settings for the various algorithms correspond to the values that the authors have recommended in their publications (Supplementary Material). For the reference method, Bimax, the discretization threshold has been set to $\underline{e} + (\bar{e} - \underline{e})/2$ where $\underline{e}$ and $\bar{e}$ represent the minimum respectively maximum expression values in the data matrix.

As the number of generated biclusters varies strongly among the considered methods, a filtering procedure, similar to Tanay *et al*. (2002) and Ihmels *et al*. (2002), has been applied to the output of the algorithms to provide a common basis for the comparison. The filtering procedure adopted here follows a greedy approach: in each step, the largest of the remaining biclusters is chosen that has less than o percent of its cells in common with any previously selected bicluster; the algorithm stops if either q biclusters have been selected or none of the remaining ones fulfills the selection criterion. For the synthetic datasets, q equals the number of optimal biclusters, which is known beforehand, and for the real datasets, q is set to 100; in both cases, a maximum overlap of $o = 0.25$ is considered.

RESULTS

Synthetic data

The data derived from the aforementioned artificial model enable us to investigate the capability of the methods to recover known groupings, while at the same time further aspects like noise and regulatory complexity can be systematically studied. The datasets used in this context are kept small, i.e. $n = 100$, $m = 50$ for scenario 1 and $n = 100$, $m = 100, \ldots, 108$ for scenario 2, in order to allow a large number of numerical experiments to be performed—for a 100×100-matrix, the running-times of the selected algorithms varied between 1 and 120 s. The size of the considered datasets, though, does not restrict the generality of the results presented in the following, since the inherent structure of the data matrix, i.e. the overlap degree, is the main focus of our study.

Note that the input matrices have not been discretized beforehand, e.g. converted into binary matrices as required by the reference method Bimax. Instead, for each algorithm the corresponding preprocessing procedures have been employed as described in the relevant papers.

Effects of noise The first artificial scenario, where all biclusters are non-overlapping, serves as a basis to assess the sensitivity of the methods to noise in the data. It is to be expected that hierarchical clustering works well in such a scenario as the implanted gene groups are clearly separated in the condition dimension.

Noise is imitated by adding random values drawn from a normal distribution to each cell of the original gene expression matrix. The noise level, i.e. the standard deviation σ, is systematically increased, and for each noise value, 10 different data matrices have been generated from the original gene expression matrix E. The performance of each algorithm is averaged over these 10 input matrices. Figure 2a summarizes the performances of the considered methods with respect to constant biclusters, while Figure 2b depicts the results for the matrices where the implanted biclusters represent trends over the conditions.

In the absence of noise, ISA, Samba and Bimax are able to identify a high percentage (>90%) of implanted transcription modules; as expected, the same holds for the hierarchical clustering approach, if the number k of clusters to be generated corresponds to the actual number of implanted modules. In contrast, the scores obtained by CC and xMotif are substantially lower. In the case of constant biclusters, this phenomenon can be explained by the fact that the largest biclusters found by these two methods mainly contain 0-cells, i.e. the algorithms do not focus on changes in gene expression, but consider the similarity of the selected cells as the only clustering criterion. This problem has been discussed in detail in Cheng and Church (2000). For the specific scenario with the particular type of additive biclusters considered here, CC tends

to find large groups of genes extending over a few columns only, which owes to the used greedy heuristic; theoretically, the implanted biclusters achieve the optimal mean residue score. Since xMotif is mainly designed to find biclusters with coherent row values, the underlying bicluster problem formulation is not well suited for the second bicluster type. A similar argument applies to OPSM which seeks clear trends of up- or down-regulation and cannot be expected to perform well in the scenarios with constant biclusters. The high average bicluster relevance in Figure 2a is rather an artifact of the implementation used in this paper which keeps the order of the columns when identical expression values are present; however, as soon as noise is added, this artificial order is destroyed, which in turn leads to considerably lower gene match scores. In contrast, biclusters following an additive model with respect to the condition dimension represent optimal order-preserving submatrices. In this setting, the correspondence between the implanted biclusters and those found by OPSM is ∼50%, cf. Figure 2b. A potential reason for the unexpectedly low scores is the way the heuristic algorithm works: per number of columns, only a single bicluster is considered—however, the implanted biclusters all extend over the same number of columns.

Concerning the influence of noise, ISA is only marginally affected by either type of noise and still recovers >90% of all implanted modules even for high noise levels. The same holds for Bimax in the constant bicluster case, but for the other bicluster type a substantial decrease in the relevance score can be observed in Figure 2b. This reveals a potential problem with discretization approaches: as noise blurs the differences between background and biclusters, many small submatrices emerge that not necessarily are meaningful. With HCL, noise has no observable effects in the constant bicluster scenarios, while for the second bicluster type increasing noise leads to a decrease in performance. The latter observation attributable to the fact that background and biclusters are not that clearly separated in the datasets with biclusters exhibiting trends. Samba seems to be sensitive to noise in the constant bicluster case as the average gene match scores decrease by 40–50% for a medium noise level; still, the scores are significantly larger than for CC and xMotif. In the case of additive biclusters, noise has only little effect on the performance of Samba. Concerning OPSM, noise affects the outcome; the scores slightly decrease. Remarkably, the performance of CC on the constant bicluster matrices appears to improve with increasing noise. This phenomenon, though, is again a result of the adopted algorithmic strategy, cf. Cheng and Church (2000): the largest biclusters may mainly cover the background, i.e. 0-cells. With noise, the biclusters found in the matrix background tend to be smaller, and this results in an improved gene match score; further evidence is provided in the supplementary material.

Regulatory complexity The focus of the second artificial scenario is to study the behavior of the chosen algorithms with respect to increased regulatory complexity. Here, a single gene may be activated by a set of transcription factors, and accordingly the implanted transcription modules may overlap. This setting is expected to reveal the advantages of the biclustering approach over traditional clustering methods such as hierarchical clustering.

Figure 2c (constant biclusters) as well as Figure 2d (additive biclusters) depict the results for different overlap degrees in the absence of noise, cf. the description of the datasets in Section ''Validation using synthetic data'' on page 1125. The only method

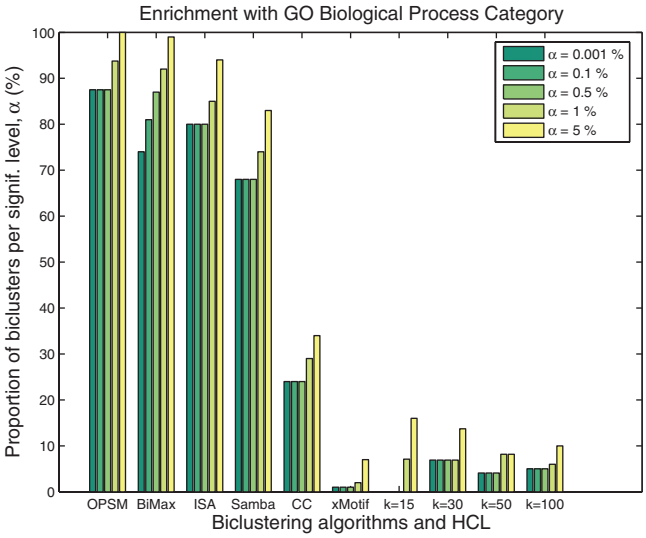

Fig. 3. Proportion of biclusters significantly enriched by any GO Biological Process category (*S. cerevisiae*) for the six selected biclustering methods as well as for hierarchical clustering with $k \in \{15, 30, 50, 100\}$. The columns are grouped method-wise, and different bars within a group represent the results obtained for five different significance levels α.

that fully recovers all hidden modules in the data matrix is—by design—the reference method, Bimax. Among the remaining methods, Samba provides the best performance: most of the biclusters found (>90%) represent hidden modules[2]; however, not all implanted modules are recovered. While OPSM is not significantly affected by the overlap degree (only the non-constant bicluster datasets have been considered as OPSM cannot handle identical expression values), ISA appears to be more sensitive to increased regulatory complexity, especially with the second bicluster type. An explanation for this is the normalization step in the first preprocessing step of the algorithm. With increasing overlap, the expression value range after normalization becomes narrower. As a result, the differences between unchanged and up- or down-regulated expression values blur and are more difficult to separate based on the gene and chip threshold parameters t_g, t_c. These parameters have a strong impact on the performance as shown in the Supplementary Material. As to CC, the performance increases with larger overlaps degrees, but the gene match scores are still lower than the ones by Bimax, Samba and ISA; again, this owes to the fact that the number of background cells diminishes with larger overlaps. xMotif shows the same behavior on the data matrices with constant biclusters. Comparing the biclustering methods with HCL, one can observe that already a minimal overlap causes a large decrease in the performance of HCL—even if the optimal number of clusters is used. The reason is that clusters obtained by HCL form a partition of genes, i.e. are non-overlapping, and this implies that not every planted transcription module can be possibly recovered.

Real data

Any artificial scenario inevitably is biased regarding the underlying model and only reflects certain aspects of biological reality.

[2]As to the outlier in Figure 2d at overlap degree 7, repeated applications of Samba on the same matrix yielded similar scores.

Table 1. Biological relevance of biclusters with respect to a metabolic pathway map (MPM) for *A. thaliana* and a protein–protein interaction network (PPI) for *S. cerevisiae*

Method	Proportion of disconnected gene pairs				Average shortest distance in the graph			
	Smaller		Greater		Smaller		Greater	
	MPM	PPI	MPM	PPI	MPM	PPI	MPM	PPI
Bimax	58.9	14.0	19.5	64.0	85.3	58.0	3.4	16.0
CC	70.0	52.0	15.0	26.0	70.0	42.0	15.0	34.0
OPSM	42.8	18.8	28.6	50.0	92.9	56.3	0.0	43.8
Samba	41.6	0.0	37.5	100.0	75.6	25.6	13.1	46.2
xMotif	49.0	2.0	17.0	92.0	84.0	4.0	3.0	72.0
ISA	25.0	58.0	25.0	22.0	50.0	70.0	25.0	22.0

For each bicluster, a *Z*-test is carried out to check whether its score is significantly smaller or greater than the expected value for random gene groups; the table gives for each method the proportion of biclusters with statistically significant scores (significance level $\alpha = 10^{-3}$). The results for HCL are omitted as all scores equal 0%.

Therefore, the algorithms are tested in the following on real datasets, normalized using mean centering, and the biological relevance of the obtained biclusters is evaluated with respect to GO annotations, metabolic pathway maps and protein–protein interaction data.

Functional enrichment The histogram in Figure 3 reflects for each method the proportion of biclusters for which one or several GO categories are overrepresented—at different levels of significance. Best results are obtained by OPSM. Given that this approach only returns a small number of biclusters, here 12 in comparison to 100 with the other methods, it delivers gene groups that are highly enriched with the GO Biological Process category. This result is insofar interesting as the effect of the noise observed in the artificial setting does not seem to be a problem with the considered real dataset. Bimax, ISA and Samba also provide a high portion of functionally enriched biclusters, with a slight advantage of Bimax and ISA (>90% at a significance level of 5%) over Samba (>80% at a significance level of 5%). In contrast, the scores for CC are considerably lower (∼30%) due to the same problem as discussed in the previous section. Cheng and Church (2000) mention that the first few biclusters should probably be discarded, but the practical issue remains that it is not clear which biclusters are meaningful and should be considered for further analysis.

Except for xMotif, though, all biclustering methods achieve higher scores than HCL with different values for k, the number of clusters to be sought. This can be explained in terms of the dataset used: Since it refers to different types of stresses, it is likely that local, stress-dependent expression patterns emerge that are hard to find by traditional clustering techniques. This hypothesis is also supported by the fact that most functionally enriched biclusters only contain one or two overrepresented GO categories and that there is no clear tendency towards any of the categories.

Comparison to metabolic and protein networks Under the assumption that the structure of a metabolic pathway map, respectively, a protein–protein interaction network is somehow reflected in the gene expression data, the degree of connectedness of the genes associated with a bicluster can be used to assess its biological relevance. In particular, one may expect that both the number of disconnected gene pairs and the average shortest distance between connected gene pairs tend to be smaller for the biclusters found than for random gene groups.

Table 1 shows that this holds for the dataset and the metabolic pathway map used for *A.thaliana*. If there exists a path between two genes of a bicluster in the metabolic graph, then with high probability the distance between these genes is significantly smaller than the average shortest distance between randomly chosen gene pairs. Although for most methods, the biclusters are better connected than random gene groups, the differences to the random case are not as striking as for the average gene pair distance. This indicates that combining gene expression data with pathway maps within a biclustering framework can be useful to focus on specific gene groups. Note that also hierarchical clustering with $k \in \{15, 30, 50, 100\}$ has been applied to these expression data; however, a single cluster usually contains almost all the genes of the dataset, while the remaining clusters comprise only few genes. Accordingly, no significant differences to random clusters can be observed.

The results for the corresponding comparison for the protein–protein interaction, though, are ambiguous, cf. Table 1. As to the degree of disconnectedness, there is no clear tendency in the data which can be attributed to the fact that not all possible protein pairs have been tested for interaction. Focusing on connected gene pairs only, ISA and Bimax seem to mostly generate gene groups that have a low average distance within the protein network in comparison to random gene sets; for xMotif, the numbers suggest the opposite. Overall, the differences between the biclustering methods demonstrate that special care is necessary when integrating gene expression and protein interaction data: not only the incompleteness of the data needs to be taken into consideration, but also the confidence in the measurements has to be accounted for, see, e.g. Gilchrist *et al*. (2004).

CONCLUSIONS

The present study compares five prominent biclusterings methods with respect to their capability of identifying groups of (locally) co-expressed genes; hierarchical clustering and a baseline biclustering algorithm, Bimax, proposed in this paper serve as a reference. To this end, different synthetic gene expression data sets corresponding to different notions of biclusters as well as real transcription profiling data are considered. The key results are as follows:

- In general, the biclustering concept allows to identify groups of genes that cannot be found by a classical clustering approach that always operates on all experimental conditions. On the one hand, this is supported by the observation that with increased regulatory complexity the ability of hierarchical clustering to recover the implanted transcription modules in an artificial

scenario decreases substantially. On the other hand, on real data the groups outputted by hierarchical clustering for different similarity measures and parameters do not exhibit any significant enrichment according to GO annotations and metabolic pathway information. In contrast, most biclustering methods under consideration are capable of dealing with overlapping transcription modules and generate functionally enriched clusters.

- There are significant performance differences among the five biclustering methods. On the real datasets, ISA, Samba and OPSM provide similarly good results: a large portion of the resulting biclusters is functionally enriched and indicates a strong correspondence with known pathways. In the context of the synthetic scenarios, Samba is slightly more robust regarding increased regulatory complexity, but also more sensitive regarding noise than ISA. While Samba and ISA can be used to find multiple biclusters with both constant and coherently increasing values, OPSM is mainly tailored to identify a single bicluster of the latter type. Proposed extensions of the OPSM approach such as Liu and Wang (2003) may resolve these issues. The remaining two algorithms, CC and xMotif, both tend to generate large biclusters that often represent gene groups with unchanged expression levels and therefore not necessarily contain interesting patterns in terms of, e.g. co-regulation. Accordingly, the scores for CC and xMotif are significantly lower than that for the other biclustering methods under consideration.
- The Bimax baseline algorithm presented in this paper achieves similar scores as the best performing biclustering techniques in this study. This may be explained by the rather global evaluation approach pursued here, and a more specific analysis may lead to different results. Nevertheless, the reference method can be useful as a preprocessing step by which potentially relevant biclusters may be identified; later, the chosen biclusters can be used, e.g. as an input for more accurate biclustering methods in order to speed up the processing time and to increase the bicluster quality. An advantage of Bimax is that it is capable of generating all optimal biclusters, given the underlying binary data model.

ACKNOWLEDGEMENTS

A.P., S.B., P.Z. and A.W. have been supported by the SEP program at ETH Zurich under the Project TH-8/02-2.

Conflict of Interest: none declared.

REFERENCES

Alexe,G., Alexe,S., Crama,Y., Foldes,S., Hammer,P.L. and Simeone,B. (2002) Consensus algorithms for the generation of all maximal bicliques. *Technical Report TF-DIMACS-2002-52*.

Azuaje,F. (2002) A cluster validity framework for genome expression data. *Bioinformatics*, **18**, 319–320.

Ben-Dor,A., Chor,B., Karp,R. and Yakhini,Z. (2002) Discovering local structure in gene expression data: the order-preserving sub-matrix problem. *In Proceedings of the 6th Annual International Conference on Computational Biology*, ACM Press, New York, NY, USA, pp. 49–57.

Bergmann,S. *et al.* (2003) Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **67**, 031902.

Berriz,G.F. *et al.* (2003) Characterizing gene sets with FuncAssociate. *BioInformatics*, **19**, 2502–2504.

Cheng,Y. and Church,G. (2000) Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* pp. 93–103.

Datta,S. and Datta,S. (2003) Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, **19**, 459–466.

Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Getz,G. *et al.* (2000) Coupled two-way clustering analysis of gene microarray data. *Proc. Natl Acad. Sci. USA*, **97**, 12079–12804.

Getz,G. *et al.* (2003) Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data. *Bioinformatics*, **19**, 1079–1089.

Gilchrist,M.A. *et al.* (2004) A statistical framework for combining and interpreting proteomic datasets. *Bioinformatics*, **20**, 689–700.

Gasch,A.P. *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

Gat-Viks,I. *et al.* (2003) Scoring clustering solutions by their biological relevance. *Bioinformatics*, **19**, 2381–2389.

Halkidi,M. *et al.* (2001) On clustering validation techniques. *J. Intell. Inform. Syst.*, **17**, 107–145.

Handl,J. *et al.* (2005) Computational cluster validation in post-genomic data analysis. *Bioinformatics*, **21**, 3201–3212.

Hartigan,J.A. and Wong,M.A. (1979) A *k*-means clustering algorithm. *Appl. Stat.*, **28**, 100–108.

Ideker,T. *et al.* (2002) Discovering regulatory and signaling circuits in molecular interaction networks. *Bioinformatics*, **18** (Suppl. 1.), S233–S240.

Ihmels,J. *et al.* (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Genet.*, **31**, 370–377.

Ihmels,J. *et al.* (2004) Defining transcription modules using large-scale gene expression data. *Bioinformatics*, **20**, 1993–2003.

Kerr,M.K. and Churchill,G.A. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl Acad. Sci. USA*, **98**, 8961–8965.

Kluger,Y. *et al.* (2003) Spectral biclustering of microarray cancer data: co-clustering genes and conditions. *Genome Res.*, **13**, 703–716.

Liu,J. and Wang,W. (2003) OP-clusters: clustering by tendency in high dimensional space. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM),* IEEE Computer Society Press, Los Alamitos, CA, USA, pp. 187–194.

Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **1**, 24–45.

Murali,T.M. and Kasif,S. (2003) Extracting conserved gene expression motifs from gene expression data. *Pac. Symp. Biocomput.*, **8**, 77–88.

Sharan,R. *et al.* (2003) CLICK and EXPANDER: a system for clustering and visualizing gene expression data. *Bioinformatics*, **14**, 1787–1799.

Sokal,R.R. and Michener,C.D. (1958) A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.*, **38**, 1409–1438.

Salwinski,L. *et al.* (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.

Tanay,A. *et al.* (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18** (Suppl. 1), S136–S144.

Tanay,A. *et al.* (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl Acad. Sci. USA*, **101**, 2981–2986.

Westfall,P.H. and Young,S.S. (1993) *Resampling-Based Multiple Testing*. Wiley, New York.

Wille,A. *et al.* (2004) Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biol.*, **5**, R92.

Yang,J., Wang,W., Wang,H. and Yu,P.S. (2002) Delta-clusters: capturing subspace correlation in a large data set. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE).* IEEE Computer Society Press, Los Alamitos, CA. pp. 517–528.

Yang,J., Wang,H., Wang,W. and Yu,P.S. (2003) Enhanced biclustering on expression data. In *Third IEEE Sympesium on Bioinformatics and Bioengineering, BIBE 2003,* pp. 321–327.

Yeung,K.Y. *et al.* (2001) Validating clustering for gene expression data. *Bioinformatics*, **17**, 309–318.

Zien,A. *et al.* (2000) Analysis of gene expression data with pathway scores. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 407–417.