BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 15 2007, pages 1927–1935
doi:10.1093/bioinformatics/btm276

Gene expression

Mining co-regulated gene profiles for the detection of functional associations in gene expression data

Attila Gyenesei[1,3,†], Ulrich Wagner[2,†,*], Simon Barkow-Oesterreicher[2], Etzard Stolte[1] and Ralph Schlapbach[2]

[1]Knowledge and Data Analysis, Unilever Research Vlaardingen, 3130 AC Vlaardingen, The Netherlands
[2]Functional Genomics Center Zürich, Uni ETH Zürich, CH-8057 Zürich, Switzerland
[3]Bioinformatics Unit, GenoSyst Ltd, Itäinen Pitkäkatu 4B, 20520 Turku, Finland

Received on February 16, 2007; revised on April 25, 2007; accepted on May 16, 2007

Advance Access publication May 30, 2007

Associate Editor: Olga Tryanskaya

ABSTRACT

Motivation: Association pattern discovery (APD) methods have been successfully applied to gene expression data. They find groups of co-regulated genes in which the genes are either up- or down-regulated throughout the identified conditions. These methods, however, fail to identify similarly expressed genes whose expressions change between up- and down-regulation from one condition to another. In order to discover these hidden patterns, we propose the concept of mining co-regulated gene profiles. Co-regulated gene profiles contain two gene sets such that genes within the same set behave identically (up or down) while genes from different sets display contrary behavior. To reduce and group the large number of similar resulting patterns, we propose a new similarity measure that can be applied together with hierarchical clustering methods.

Results: We tested our proposed method on two well-known yeast microarray data sets. Our implementation mined the data effectively and discovered patterns of co-regulated genes that are hidden to traditional APD methods. The high content of biologically relevant information in these patterns is demonstrated by the significant enrichment of co-regulated genes with similar functions. Our experimental results show that the Mining Attribute Profile (MAP) method is an efficient tool for the analysis of gene expression data and competitive with bi-clustering techniques.

Contact: ulrich.wagner@fgcz.ethz.ch

Supplementary information: Supplementary data and an executable demo program of the MAP implementation are freely available at http://www.fgcz.ch/publications/map

1 INTRODUCTION

The application of mRNA gene expression microarrays has proven to be an invaluable tool for the elucidation of mechanisms of diverse biological processes at the molecular level. In a microarray experiment, several thousands of genes are investigated in parallel. Mainly due to the high costs of the microarrays, gene expression studies are normally carried out with a rather limited set of conditions and repetitions, featuring an experimental design that focuses on a few very specific research questions. With time, however, collecting microarray data sets brings in a new dimension into gene expression data analysis: the investigation of a large set of genes in a large set of experimental conditions. Analyzing such data is not trivial and requires sophisticated data mining solutions.

During the last decade, both supervised and unsupervised data mining methods have been applied to gene expression data. Supervised techniques, such as support vector machines (Brown, 2000) and artificial neural networks (Vohradsky, 2001), aim at building a robust classifier on predefined sample groups to assign any new sample to the proper group. This kind of analysis can be applied successfully to clinical diagnostics. Unsupervised methods concentrate on understanding the similarity of gene expression profiles among all samples by grouping similarly expressed genes together with the idea that genes with similar expression profiles might share common regulatory mechanisms and functions. There are two main types of unsupervised data analysis, namely dimensionality reduction, e.g. principal component analysis (Raychaudhuri, 2000) and singular value decomposition (Alter *et al.*, 2000) and clustering such as k-means (Tavazoie *et al.*, 1999), hierarchical clustering (Eisen *et al.*, 1998) and self-organizing maps (Tamayo *et al.*, 1999). For good overviews of the above techniques, we refer to, e.g. Leung and Cavalieri, 2003 and Quackenbush, 2001.

More recently, bi-clustering (Ben-Dor *et al.*, 2003; Cheng and Church, 2000; Ihmels *et al.*, 2002; Prelic *et al.*, 2005; Xu *et al.*, 2006) and association pattern discovery (APD) methods (Carmona-Saez *et al.*, 2006; Creighton and Hanash, 2003; Georgii *et al.*, 2005) have been adapted to find patterns of co-regulated genes. In contrast to the described unsupervised and supervised techniques, these methods are able to discover co-regulated genes not only over the full set but also within and among subsets of conditions (samples). Moreover, each gene and each condition can occur in more than one cluster/pattern. While the idea of bi-clustering comes from the area of traditional clustering, namely to apply a similarity measure to calculate the correlation between cluster members, APD methods are inherited from the area of frequent itemset and association rule mining.

*To whom correspondence should be addressed.
†The authors wish it to be known that in their opinion, the first two authors should be regarded as joint First Authors.

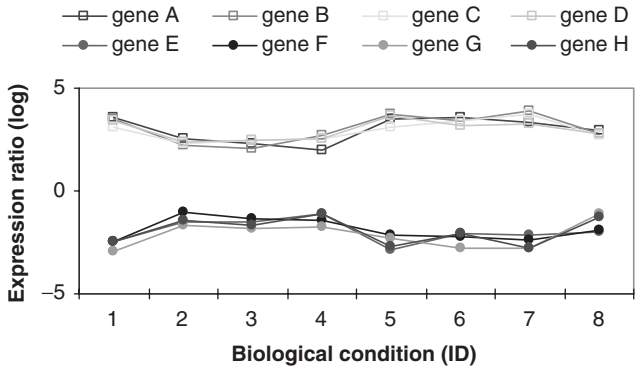


Fig. 1. A sample expression pattern that can be discovered by APD methods. All of the genes in the pattern are either up- or down-regulated throughout the identified conditions. Co-regulated genes varying between up- and down-regulations are not taken into consideration. For details, see Supplementary Material.

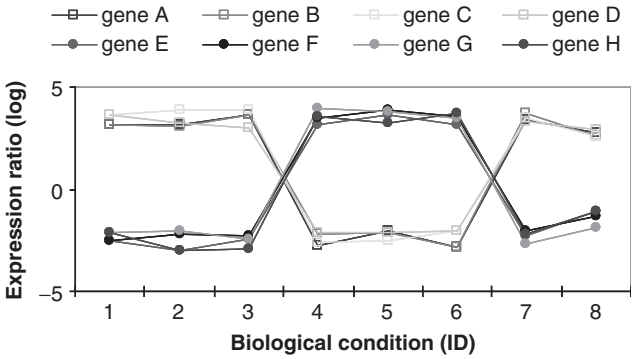


Fig. 2. A synthetic gene-expression data set. Traditional APD methods are unable to identify the co-regulation neither between the eight genes nor between the four ones, *A*, *B*, *C*, *D* or *E*, *F*, *G*, *H* for all biological conditions. However, the absolute correlation between the expression ratios of the eight genes is very close to 1. A relationship between the eight genes can be summarized as {**gene *A*, gene *B*, gene *C*, gene *D* (−) gene *E*, gene *F*, gene *G*, gene *H***}, which is the only maximal (and closed) ***co-regulated gene profile*** of the synthetic data. For details, see Supplementary Material.

Although the main focus of this article is to show the improvement that our method represents in relation to traditional APD methods, we also discuss the relationship between our method and bi-clustering techniques since they produce similar results.

APD methods can describe associations of differentially expressed genes. Relationships discovered by these means are represented in the form of expression patterns and association rules. Since association rules are generated from expression patterns as a post-processing step, most of the algorithmic research is concentrated on mining expression patterns. Such patterns are composed of groups of genes that are always up-regulated or always down-regulated throughout the identified conditions. A sample expression pattern is shown in Figure 1, where genes *A*, *B*, *C* and *D* are always up-regulated, *E*, *F*, *G* and *H* are always down-regulated for eight biological conditions.

In biological terms, restricting the search for relationships, in which the individual genes are always up- or down-regulated throughout the conditions does not make sense. In a network or a pathway of genes, an inhibition of expression of gene *A* by the expression of gene *E*, could mean an up-regulation of gene *E* when gene *A* is down-regulated (Fig. 2). Such important causal relationships of gene regulation events will not be taken into account by traditional APD methods that have so far been applied to gene expression data analysis.

In this article, we tackle the above limitation by introducing the concept of mining co-regulated gene profiles. Co-regulated gene profiles contain two sets such that genes within the same set behave identically (up or down) while genes from different sets display contrary behavior (for precise definitions, see Methods Section). Such behaviors are calculated directly from the original gene expression data without the need of data transformation, as proposed in Ji and Tan (2004). Moreover, our method is able to discover inverse co-regulations not only between a single gene and a set of genes but between any sets of genes, as shown in Figure 2. The proposed algorithm is based on our previously developed MAP (Mining Attribute Profile) method that has been shown to be able to discover hidden relationships between the whole attributes based on their 'changing tendency' in a condition-based manner (Gyenesei *et al.*, 2006). Therefore, applying the MAP method to gene expression data allows for the identification of genes whose expression follows the same pattern in response to different biological conditions.

One important factor in microarray data is inherent noise. This limits the usefulness of APD methods as it leads to the generation of many redundant and similar patterns. While redundant patterns can be discarded by mining only the closed patterns (Pasquier, 1999), the number of similar patterns is only insufficiently reduced. To address this problem, we propose a new similarity measure that can be applied together with hierarchical clustering and leads to grouped similar patterns. Experimental results show that previously hidden co-regulated genes with high correlation can be discovered by the proposed methods.

The remainder of the article is organized as follows: first we briefly review the concept of traditional association pattern discovery before outlining its limitations. Then the definition of co-regulated gene profiles is introduced and the main steps of the proposed algorithm are summarized. This is followed by the discussion of how similar patterns are grouped together. Section 2 concludes with a description of the methods used to measure the biological relevance of the discovered profiles. The experimental validation of the MAP algorithm is reported in Section 3 based on the application of the method to two yeast expression data. Finally, discussion and further outlooks are presented in Section 4.

2 METHODS

2.1 Association pattern discovery

The problem of APD originates from market basket analysis, which aims at finding interesting relationships hidden in large data sets. Such relationships can be represented in the form of *frequent itemsets*

and *association rules*. APD is a two-step process: first the frequent itemsets are discovered from which, as a second step, the association rules are generated. For the precise formulation of the problem, see Agrawal and Srikant (1994).

Since its introduction, APD has been successfully applied not only to market basket analysis but also to many other industrial and scientific research problems, and recently to gene expression data. In this context, an itemset represents a group of genes with their expressions being increased, decreased or not changed under a specific set of conditions. Such itemsets can be termed *expression patterns*. As an example, an expression pattern in gene expression data {gene $A\uparrow$, gene $B\uparrow$, gene $C\downarrow$} indicates that gene A and B are up-regulated and gene C is down-regulated in a sufficient number of conditions. Similarly, a strong association rule {gene $C\downarrow$} => {gene $A\uparrow$, gene $B\uparrow$} indicates that if gene C is down-regulated, then both genes A and B are up-regulated for a given confidence threshold.

2.2 Limitations of traditional APD methods applied to gene expression data

The limitations of existing APD methods can be derived from the origin of the research problem. Basket analysis aims at understanding the behavior of customers based on their shopping baskets, in which an item is purchased or not. Consequently, traditional APD algorithms were developed for databases containing only binary attributes. When these algorithms are applied to gene expression data containing continuous expression values, an additional preprocessing step is employed to transform the continuous attribute domains into categorical ones (discretization) and the obtained categorical domains into binary attributes (binarization). The problem with such preprocessing is that the discovered expression patterns no longer reflect the associations between the (whole) set of genes but the relations between their binned independent expression values (such as, up-regulated and down-regulated). Therefore, APD algorithms can discover only those expression patterns in which all genes are either up- or down-regulated throughout the identified conditions (Fig. 1). They are even unable to discover the co-regulation between genes having identical expression values if they are changing between up- and down-regulation from one condition to another (Fig. 2). This limitation could be overcome by identifying the co-regulated genes based on their expression behaviors instead of their strict up- and down-regulation.

In the next sections, we introduce the concept of mining co-regulated gene profiles in which gene profiles are defined and mined based only on their changing behaviors. This approach allows for the identification of genes whose expression profiles follow the same patterns in response to different biological conditions. Applying this concept, previously unknown co-regulated genes can be discovered that remain hidden to traditional approaches.

2.3 Co-regulated gene profiles

Let $E=[e_{g,c}]_{n\times m}$ be the normalized gene expression matrix over a set of n genes and m microarray experiments (biological conditions). A matrix element $e_{i,u}$ denotes the log-fold expression change of gene $g_i \in G$, $G=\{g_1,g_2,\ldots,g_n\}$, at biological condition $c_u \in C$, $C=\{c_1,c_2,\ldots,c_m\}$. Let δ be a user-defined log-fold change threshold. A gene g_i is said to be *up-regulated* at experimental condition c_u if its log-fold expression change $e_{i,u}$ is not less than the defined threshold δ. Similarly, g_i is *down-regulated* at condition c_u if $e_{i,u}$ is not higher than $-\delta$.

Let g_i and g_j be two genes. We say that g_i and g_j have *identical behavior* at experimental condition c_u if their log-fold expressions are either higher than or equal to fold-change threshold δ or less than or equal to $-\delta$:

$$\forall g_i,g_j \in G,\forall c_u \in C : (e_{i,u} \geq \delta \,\&\, e_{j,u} \geq \delta) \text{ or } (e_{i,u} \leq -\delta \,\&\, e_{j,u} \leq -\delta).$$

In other words, two genes behave identically for a certain condition if they have the same (up or down) regulation.

Similarly, the *contrary behavior* between genes g_i and g_j at experimental condition c_u is defined as follows:

$$\forall g_i,g_j \in G,\forall c_u \in C : (e_{i,u} \geq \delta \,\&\, e_{j,u} \leq -\delta) \text{ or } (e_{i,u} \leq -\delta \,\&\, e_{j,u} \geq \delta).$$

Therefore, two genes behave contrary for a certain biological condition if they have different regulation.

Let $U \subseteq C$ be a set of experimental conditions and I, $J \subseteq G$ be two contrary behaved sets of genes over condition set U, where genes in both sets behave identically. The formula $\{I\ (-)\ J\}$ is called a *co-regulated gene profile* of genes $I \cup J$ for condition set U.

Consider again the synthetic data given in Figure 2 to demonstrate the above definitions. There are two sets of genes, namely {gene A, gene B, gene C, gene D} and {gene E, gene F, gene G, gene H}, that contain genes with identical behavior; i.e. whenever the expression level of one of the genes is affected in a specific way (up- or down-regulated), the expression level of the other three genes are affected in the same way in all conditions. Moreover, the behavior of the two sets of genes is inverted. Therefore, the co-regulated gene profile between the eight genes can be formulated as {gene A, gene B, gene C, gene D (−) gene E, gene F, gene G, gene H}.

Similarly to APD methods, the research task of co-regulated gene profile discovery is to find all profiles that exist in at least as many number of experimental conditions as a user-defined *minimum support threshold* σ. Profiles that satisfy the support requirement are called *frequent co-regulated gene profiles*.

2.4 Algorithm of mining co-regulated gene profiles

We have developed an efficient algorithm to discover co-regulated gene profiles in large gene expression data. The algorithm can be characterized as a depth-first search, divide-and-conquer algorithm. We have chosen this type of searching strategy in order to reduce the number of database scans and avoid the costly set-containment-test operation that can be the case when applying a breadth-first search strategy. The mining part is carried out in two steps in which the first step constructs a compact data structure called Gene Profile tree (or GP-tree), and the second step extracts the frequent co-regulated gene profiles directly from the GP-tree structure.

Due to space constraints, we are not able to present the precise algorithm and have to refer to Gyenesei *et al.* (2006) for more details. For an illustrative example, see Supplementary Material. Here, we just summarize the main ideas of the two steps as follows:

(1) *Constructing a GP-tree*. A GP-tree is constructed by reading the expression data condition by condition and mapping each condition onto a path in the GP-tree. A path compression occurs when two or more conditions have the same gene profile starting from the first gene in the tree. More overlapped paths result in a more compressed data set and a smaller tree. As a consequence, the mining algorithm needs less time to extract the frequent co-regulated profiles from the GP-tree structure.

(2) *Mining co-regulated gene profiles using the GP-tree*. The developed mining algorithm generates co-regulated gene profiles from the constructed GP-tree by exploring the tree in a top–down and recursive manner. It splits the problem into sub-problems by decomposing the GP-tree into disjoint sub-GP-trees, and then calls the recursion again with the sub-trees. If the constructed sub-GP-tree has only a single branch, then all co-regulated gene profiles are enumerated directly from the single branch.

2.5 Grouping gene profiles by similarity

Unfortunately, noise is inherent to microarray data and can significantly increase the number of discovered patterns. Most of the redundant patterns can be discarded by applying the idea of mining closed and maximal patterns (Goethals and Zaki, 2003), but still, also because of the number of significant patterns, too many patterns are reported for the users. Therefore, grouping similar patterns is an important step in our concept, as it allows biologists to get a general picture about the discovered patterns and to study the most interesting ones in more detail.

The similarity between co-regulated gene profiles can be measured as follows. Let $P_i = \{P_i^L - P_i^R\}$ and $P_j = \{P_j^L - P_j^R\}$ be two gene profiles where P_i^L, P_i^R, P_j^L and P_j^R are sets of genes. The similarity between P_i and P_j is defined by

$$s(P_i,P_j) = \max\{P_i \Delta P_j, P_i \nabla P_j\}/|P_i^L \cup P_i^R \cup P_j^L \cup P_j^R|,$$

where $P_i \triangle P_j$ denotes the number of those genes, which are in both P_i^L and P_j^L or in both P_i^R and P_j^R:

$$P_i \Delta P_j = |P_i^L \cap P_j^L| + |P_i^R \cap P_j^R|,$$

and $P_i \nabla P_j$ denotes the number of genes which are in both P_i^R and P_j^L or in both P_i^L and P_j^R as formulated by

$$P_i \nabla P_j = |P_i^L \cap P_j^R| + |P_i^R \cap P_j^L|.$$

To illustrate the calculation of the similarity measure, let $P_1 = \{1,3,5\,(-)\,2,4\}$, $P_2 = \{4\,(-)\,1,5\}$ and $P_3 = \{1,5\,(-)\,2,4\}$ be gene profiles. Then, $s(P_1, P_2) = \max\{0,2+1\}/5 = 0.6$, $s(P_1, P_3) = \max\{2+2,0\}/5 = 0.8$ and $s(P_2, P_3) = \max\{0,1+2\}/4 = 0.75$.

Having defined the similarity measure between profiles, we applied the idea of *agglomerative hierarchical clustering* to group them. Therefore, the algorithm starts with the single profiles as individual groups and, at each step, it merges the most similar pair of profile groups. The algorithm stops when only one group is left containing all of the gene profiles.

Of course, one of the key points in the clustering algorithm is the calculation of similarity between two groups if they contain more than one profiles. Our algorithm can handle the most commonly used techniques which are *single-*, *complete-* and *average-linkage*, i.e. taking the maximum, minimum or average similarity of all pair-wise similarities between two profile groups. In addition, our similarity measure is able to capture the similarities not only between profiles but between genes, simply replacing profiles by genes in the above formula. This allows biologists to represent the clustering results by well known visualization tools, such as *MapleTree* developed by Eisen's Lab (http://rana.stanford.edu), as shown in Figure 3.

2.6 Measuring the biological relevance of the discovered profiles

Finding groups of genes that are coordinately correlated throughout a set of experiments leads to the assumption that these genes are involved in common functions and/or roles. Thus, integrating a priori knowledge on functions of single genes into the analysis of these groups of genes will allow hypothesizing on a common function. Significantly over-represented functional categories such as the Gene Ontology (GO) groups (Ashburner *et al.*, 2000) or *KEGG* pathways (Kanehisa *et al.*, 2006) can therefore be determined in the gene patterns discovered using our proposed algorithm. The over-representation analysis was carried out using hypergeometric distributions as implemented in the *ermineJ* software (Lee *et al.*, 2005). The resulting *P*-values were adjusted using Bonferroni's multiple testing correction method. Overrepresentation analysis of KEGG pathway membership was carried out using a Fisher's exact test implementation in the statistical package *R*.

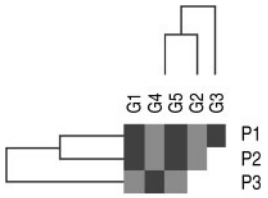


Fig. 3. Visualizing the sample profile data. The data has been clustered by profile- and gene similarities as defined in the text. The two colors (black and grey) do not represent expression ratios but behaviors between genes in a profile. For example, in the first profile, there are two sets of genes with identical behaviors, namely {1,5,3} (colored by black) and {4,2} (colored by grey), and they behave contrary to each other.

2.7 Relation to bi-clustering methods

Due to the resemblance of results, the MAP method competes with existing biclustering methods. We performed a short comparison by using the freely available Biclustering Analysis Toolbox (BiCAT) (Barkow *et al.*, 2006) with the same data sets. BicAT implements a number of common biclustering methods including: (i) Cheng and Church's algorithm (CC), which is based on a mean squared residue score (Cheng and Church, 2000); (ii) the Iterative Signature Algorithm (ISA) that searches for submatrices representing fix points (Ihmels *et al.*, 2002); (iii) the Order-preserving Submatrix Algorithm (OPSM), which tries to identify large submatrices for which the induced linear order of the columns is identical for all rows (Ben-Dor *et al.*, 2003); (iv) Bimax, an exact biclustering algorithm based on a divide-and-conquer strategy that is capable of finding all maximal bicliques in a corresponding graph-based matrix representation (Prelic *et al.*, 2005).

3 EXPERIMENTAL RESULTS

To demonstrate the usability and efficiency of the concept of mining co-regulated gene profiles, we applied it to two publicly available gene expression data sets from *Saccharomyces cerevisiae*. The first data set (referred hereafter as Yeast80) comes from Stanford University and contains information about the expression of 6221 genes in 80 different conditions including diauxic shift, mitotic cell division cycle and sporulation (for details, refer to Eisen *et al.*, 1998). The second data set (referred hereafter as Compendium) includes expression levels of 6316 genes in 300 diverse yeast mutants or in wild type yeast with different chemical treatments (Hughes *et al.*, 2000). For both data sets, we chose a cutoff of 2-fold increase or decrease to define differential expression. The properties of the two data sets are summarized in Table 1.

3.1 Mining Co-regulated gene profiles

3.1.1 MAP compared to traditional APD methods In order to discover the co-regulated gene profiles, we applied our MAP algorithm to the yeast data sets with minimum support thresholds of 10 for both numbers of genes and conditions. Using these settings, our target was to discover closed gene expression profiles in which at least 10 genes behave (respond)

Table 1. Properties of yeast data sets used for pattern mining

	Yeast80	Compendium
Number of conditions	80	300
Number of genes	6121	6316
Log base of expression ratios	2	10
Log-fold expression threshold	±1	±0.3

Table 2. Number of co-regulated gene profiles discovered by the MAP methods

	Yeast80	Compendium
Number of profiles	340	73831
Number of hidden profiles	124	7496
Number of genes in the longest profile	55	43
Average Pearson's correlation	0.64	0.81
Average Pearson's correlation of the hidden profiles	0.93	0.94
Running time (s)	3	6

The thresholds for the minimum numbers of conditions and genes in a profile were set to 10. Traditional APD methods can discover as many patterns as the difference between all of the profiles and the hidden ones.

similarly for at least 10 biological conditions. As shown, our MAP implementation is able to discover the closed patterns automatically, preventing the accumulation of many redundant patterns. In the case of the number of closed patterns being very large, maximal patterns can be gained from the closed ones as a post-processing step. For reasons of clarity of the presented example analysis, we concentrated only on the closed patterns. The implementation furthermore supports parameters to be freely changed by the users based on individual and experimental requirements. For access to the Windows executable demo program, please refer to the project web site (see Supplementary Material).

Table 2 summarizes the mining results for both data sets. The number of hidden profiles is counted simply by checking whether they exclusively satisfy the co-regulated gene profile properties. If a profile contains at least 10 genes and 10 conditions such that each of those genes is either up-regulated or down-regulated throughout at least 10 biological conditions, then it can be discovered by traditional APD methods and is therefore not a hidden profile. To verify the number of non-hidden profiles, we applied in parallel one of the most popular frequent closed pattern mining methods, ChARM (Zaki and Hsiao, 1999), which produced the same number of non-hidden patterns. Note that any kind of such algorithms would produce the same result.

To check how the genes are correlated in the discovered hidden profiles, we calculated the absolute Pearson correlation between their real expression ratios. As Table 2 shows, the average correlation of the hidden profiles is 0.93 for the Yeast80 and 0.94 for the Compendium data. These are surprisingly high values considering that 2-fold cutoff were applied before the mining process to define differential expressions (up- and down-regulation). The graphical view of the expression behavior of the longest profile for data set Yeast80 can be seen in Figure 6 and it is denoted by letter *D*. It contains 55 genes with an average correlation of 0.93. The capability of the implemented program to illustrate the behavior of all of the discovered profiles can be found in the Supplementary Material.

Table 3. Comparison of the results obtained by different mining algorithms on the Yeast80 data set

	MAP	Bimax	OPSM	ISA	CC
Running time (s)	3	40	875	413	$>10^6$
Number of clusters	340	1127	18	36	50
Maximum number of genes	55	62	5296	418	6221
Minimum number of genes	10	10	4	14	4
Maximum number of conditions	11	11	23	28	7
Minimum number of conditions	10	10	2	9	4

3.1.2 MAP compared to bi-clustering methods In order to compare the profiles obtained by MAP and bi-clusters discovered by bi-clustering techniques, we carried out a detailed analysis using the Yeast80 data set. Bimax, OPSM, ISA and CC were selected as reference bi-clustering methods for the comparison. Table 3 summarizes the mining results for the Yeast80 data set. Detailed information on the settings of the methods as well as the composition of the resulting clusters can be found in the Supplementary Material.

As Table 3 shows, OPSM, CC and ISA had much longer running times and detected much less clusters than MAP and Bimax with default settings. Moreover, clusters discovered by the three slowest methods display a much higher variability in terms of numbers of genes and conditions. Neither OPSM nor ISA was able to identify clusters that contain inversely correlated genes. Bimax could detect such clusters and they were very similar to the hidden profiles discovered by MAP. The only difference between Bimax and MAP was that clusters obtained by Bimax contained also genes with random up- and down-regulation. Such a cluster is displayed in Figure 4 where the false positive members of the cluster are indicated by an arrow.

Finally, we tested the four bi-clustering methods and our MAP algorithm using the two simple data sets that are shown in Figures 1 and 2 (for the exact data, see Supplementary Material). The first data was recognized correctly by all methods as one cluster, whereas only Bimax and MAP were able to recognize the second data as a single cluster or profile, respectively.

3.2 Biological relevance of the MAP results

The resulting patterns from our data processing were analyzed in respect to the enrichment of functional GO categories using the overrepresentation analysis functionality of the ermineJ software (Lee *et al.*, 2005). For each category, the calculated *P*-value was corrected by the Bonferroni method as presented

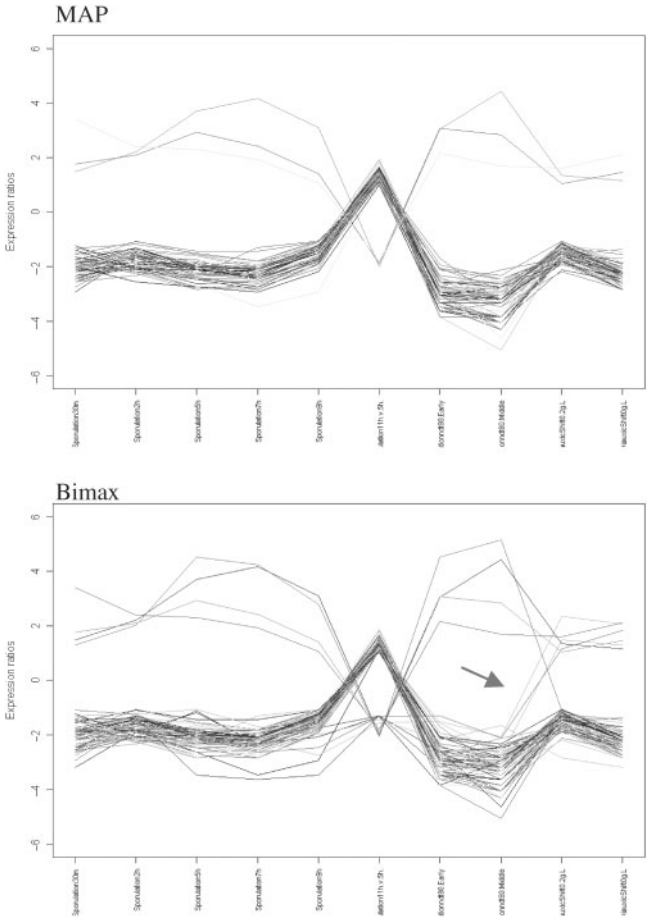


Fig. 4. The largest MAP profile and Bimax bicluster found in the Yeast80 data set. The arrow indicates the presence of additional genes in the Bimax cluster that represent false positives.

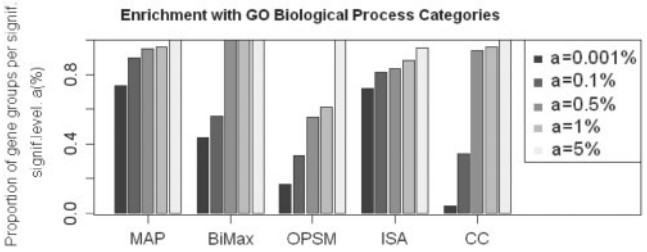


Fig. 5. Proportion of bi-clusters that show significant enrichment by any GO category (*S.cerevisiae*) for the MAP and four bi-clustering algorithms. The different bars within a group represent the results obtained for five different significance levels *a*. *P*-values were adjusted with a Bonferroni correction. GO groups of a size larger than 100 were omitted.

in Prelic *et al*. (2005). In Figure 5, the proportions of gene patterns are displayed that showed a significant enrichment of any GO category. Only those groups were taken into account that had less than 100 members. It can be seen that all of the discovered patterns show significant functional enrichment at the 5% level. When lowering the significance level to 0.01%, still 90% of the detected patterns show significant enrichment. We compared these findings to the results of the enrichment analysis of gene groups detected with the four bi-clustering methods. Figure 5 shows that our method found more gene groups with higher significance, whereas Bimax identified more bi-clusters with significant enrichment when lowering the stringency. The details of the GO analysis for each pattern identified by any of the five algorithms can be found in the Supplementary Material.

A similar analysis of the Yeast80 data set was carried out for the enrichment of *KEGG* pathways in the obtained patterns of genes. The KEGG pathways are more precise in terms of biological content than GO categories as they accurately describe the roles of genes. As a drawback, the KEGG pathway information is far from being complete, although the KEGG database contains one of the largest publicly available pathway data set for yeast. Fifty percent of the MAP patterns show enrichment of KEGG pathways at the 5% significance level, 15% at the 0.01% significance level and still more than 10% at the 0.001% significance level.

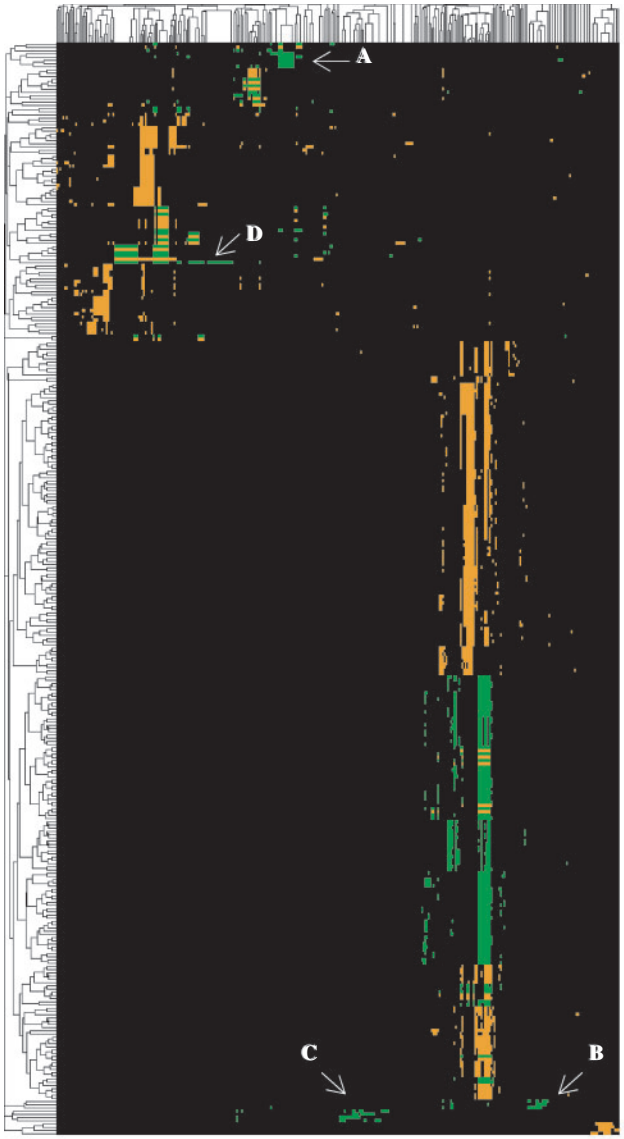
Inspecting these results more closely, we found two profiles that deserve further explanation. One of them is the longest pattern that we found with the MAP algorithm using the settings as described before. This pattern contains 55 genes, of which 21 are small ribosomal and 26 large ribosomal genes. For illustration, we have mapped the identified genes onto the KEGG pathway representation (see Supplementary Material). Another very interesting profile contains five central enzymes of the glycolysis pathway. The mapping onto the respective KEGG pathway can also be found in the Supplementary Material.

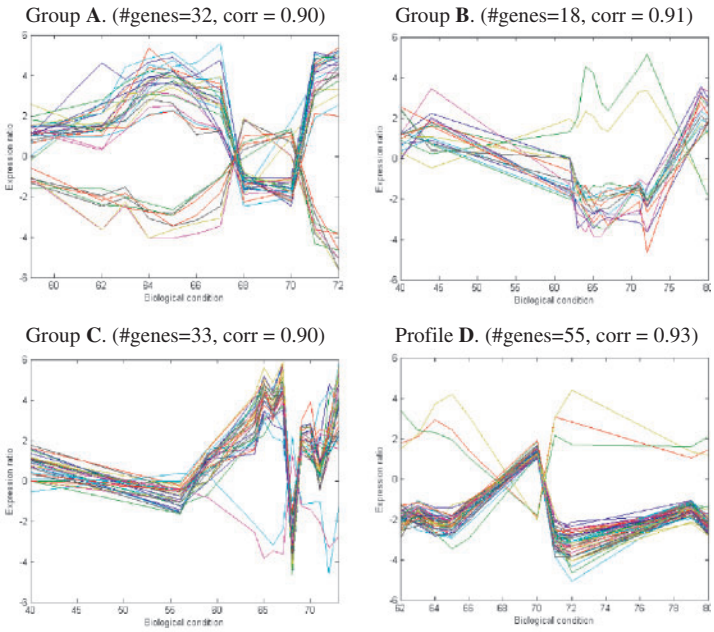
3.3 Biological relevance of hidden patterns

To reduce the noise effect of gene expression data and to group those patterns together that share similar co-regulations, we applied the average-linkage hierarchical clustering technique. The similarities between gene profiles have been calculated by the newly proposed similarity measure (see Methods section). We would like to emphasize that, to our best knowledge, no similar methods have been proposed for the traditional APD problem in the literature. During the implementation of the clustering method, our main goal was to provide visualization designed to be familiar to biologists. To accommodate this feature, the program is able to generate the required input files for the well-known visualization tool, MapleTree, which is freely downloadable from Stanford University (http://rana.stanford.edu).

Figure 6 illustrates the clustering results (left part) as well as selected hidden gene profile groups and their biological relevance (right part) for the Yeast80 data. Note that both profiles (rows) and genes (columns) have been clustered. Hidden profiles that cannot be discovered by traditional methods appear in green color.

We investigated more closely a number of pattern clusters that were not discovered by traditional APD methods. These clusters are marked with letters *A*, *B*, *C* and *D*. Next to the right upper side of the clustering picture, the gene expression





Group	Function	GO #	Overlap	Corr. P-value
A	meiosis I	GO:0007127	12	8.44E-18
	Synapsis	GO:0007129	7	6.84E-16
	meiotic recombination	GO:0007131	7	1.69E-08
B	Response to stimulus	GO:0050896	3	0.04572
	biofilm formation	GO:0042710	1	0.04942
	cell wall	GO:0005618	3	0.03758
C	spore wall assembly	GO:0042244	6	1.60E-06
	sporulation (sensu Fungi)	GO:0030437	7	2.89E-05
	cell wall (sensu Fungi)	GO:0009277	6	3.55E-04
Largest profile (**D**)	large ribosomal subunit	GO:0015934	25	6.22E-33
	small ribosomal subunit	GO:0015935	21	4.70E-29
	eukaryotic 43S preinitiation complex	GO:0016282	21	1.56E-27

Fig. 6. Grouping of the discovered profiles (rows) and their genes (columns) by the introduced similarity measure. Hidden profiles that cannot be discovered by traditional methods appear in green color. The graphical views of the expression behaviors of the three selected hidden groups (**A**, B and **C**) and the longest profile (**D**) can be seen at the right upper side of the figure. The table summarizes the three most significant GO categories for all selected patterns.

profiles of these four pattern groups are shown and the table underneath the profiles summarizes the three most significant GO categories for each profile. The patterns belonging to group *A* contain many genes that are involved in the aligning of homologous chromosomes (synapsis) and other events during meiosis. It can be hypothesized that other genes in this pattern cluster fulfill similar functions in the process of meiosis. As another interesting finding, group *B* and even more group *C* strongly suggest that these patterns contain genes that are involved in the formation of the cell wall as these patterns show highly significant enrichment of the respective GO categories. Furthermore, pattern *D*, which is strongly enriched in ribosomal proteins has already been explained in more details in section 3.2 of the results part. These groups of genes and therefore their hypothetical function would not have been found by traditional APD methods.

4 DISCUSSION

APD based methods are well established and popular techniques for the mining of transaction databases that are built up in market basket research. For the investigation of databases in the field of functional genomics, however, interest in applying APD methods is just beginning to emerge. Although important steps have been made to develop solutions

to the area of microarray data analysis, there is still room for improvement for the proposed applications.

We here presented an adaptation of the APD method that allows the mining of large microarray data sets in an efficient way. We demonstrated that the resulting patterns contain useful information for the biologists. As shown in the Yeast80 data set, all of the patterns show significant enrichment in genes that can be functionally classified into the same GO category even when using a very conservative correction for the false positive rate. We further investigated selected patterns by using a priori information from data that are less comprehensive but of higher quality, such as the KEGG pathways. This yields a nearly exact reconstruction of gene groups that are functionally related to each other, e.g. for the glycolysis pathway or for the ribosomal protein groups.

To get a better overview of the resulting patterns from MAP data mining, we used a straightforward way of representing them by applying hierarchical clustering methods and using an appropriate visualization tool. Furthermore, we proposed a new similarity measure that allows for grouping of the patterns and the respective genes within those resulting patterns.

The clustering results show that there is a considerable number of (partially) overlapping gene patterns. This can be overcome to some extent by making use of maximal instead of closed patterns, which reduces the number of patterns in the case of the Compendium data by a factor of ten (data not shown). Further improvements are presently matter of our research.

Finding patterns that overlap at least partly reflects the fuzziness of the microarray data itself. Gene expression has been shown to be intrinsically noisy as the biochemical reactions of gene expression are of stochastic or inherently random nature (Raser and O'shea, 2005). In addition to the biological source of variation, there is also a technical one, as microarray analysis involves many technical steps, each of which contributes to the variation (Coombes *et al*., 2002; Zakharin *et al*., 2005). However, it has been shown that, in a robust microarray platform, the biological variation is bigger than the technical variation.

The MAP algorithm represents a clear advancement in the applicability of APD methods to microarray data, as it produces a more comprehensive set of patterns. With the given settings for the minimum numbers of genes and conditions per pattern, the MAP algorithm detects patterns that are not detected by the traditional APD methods. Some of those patterns were enriched in genes of unique biological functionality. Therefore, such sets of genes could not be found in any pattern that was detectable by traditional methods. Moreover, a pattern obtained with the traditional APD method showing overrepresentation of the same functional categories as a pattern obtained using MAP might have been composed by a lower number of conditions.

Taken together, important pieces of information that would have been lost when mining microarray data with traditional APD methods can be revealed using the MAP algorithm.

Although the main focus of this article is to show the improvement that MAP represents in relation to traditional APD methods, we briefly discuss the relationship between MAP and bi-clustering techniques, as their results are of similar nature.

Bi-clustering techniques overcome a flaw in traditional clustering methods for gene expression data by allowing that each gene and each condition can occur in more than one cluster (Madeira and Oliveira, 2004). In addition to this, a bi-cluster can consist of a subset of genes and a subset of conditions. Profiles discovered by our MAP method also have these properties. Like bi-clustering, MAP (and APD methods in general) are very useful exploratory methods since they allow for the detection of unexpected results. At the same time, this represents a weakness as neither bi-clustering nor APD methods are statistical methods in a strict sense. Typically, no model assumptions are made, the significance of the results cannot be calculated (unless using resampling techniques) and no false positive rates can be determined. Furthermore, both methods show an enormous redundancy because of the partial overlapping of genes and conditions.

For the Yeast80 data set, we showed in the discovered patterns that MAP outmatches all tested bi-clustering techniques in terms of speed and biological significance (CC, ISA and OPSM), which can be interpreted as a metric of accuracy. Only Bimax was able to find bi-clusters that correspond to the hidden profiles detected by MAP. In general, clusters discovered by Bimax were very similar to profiles identified by MAP. However, Bimax includes obvious false positive genes into clusters, which can result in mild or disastrous errors.

To sum it up, this first non-exhaustive comparison between MAP as an improved APD method and bi-clustering methods indicates that MAP represents a competitive method. We are presently working on a more comprehensive and systematic comparison, including more data sets and different parameter settings for the above and additional methods.

ACKNOWLEDGEMENTS

The authors thank Sarah Rodgers for fruitful discussions regarding the MAP algorithm and Katalin Fülöp and Mike Scott for helpful advices and comments on the manuscript. A. Gy. is supported by a Marie Curie fellowship.

Conflict of Interest: none declared.

REFERENCES

Agrawal,R. and Srikant,R. (1994) Fast algorithms for mining association rules. In *20th VLDB Conference*.

Alter,O. *et al*. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, **97**, 10101–10106.

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

Barkow,S. *et al*. (2006) BicAT: a biclustering analysis toolbox. *Bioinformatics*, **22**, 1282–1283.

Ben-Dor,A. *et al*. (2003) Discovering local structure in gene expression data: the order-preserving submatrix problem. *J. Comput. Biol*., **10**, 373–384.

Brown,M.P. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Carmona-Saez,P. *et al*. (2006) Integrated analysis of gene expression by Association Rules Discovery. *BMC Bioinformatics*, **7**, 54.

Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **8**, 93–103.

Creighton,C and Hanash,S. (2003) Mining gene expression databases for association rules. *Bioinformatics*, **19**, 79.

Coombes,K.R. *et al.* (2002) Identifying and quantifying sources of variation in microarray data using high-density cDNA membrane arrays. *J. Comput. Biol.*, **9**, 655–669.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Georgii,E. *et al.* (2005) Analyzing microarray data using quantitative association rules. *Bioinformatics*, **21**, ii123–ii129.

Goethals,B. and Zaki,M.J. (2003) FIMI'03: workshop on frequent itemset mining implementations. 19th December 2003, Melbourne, Florida, USA.

Gyenesei,A. *et al.* (2006) Frequent pattern discovery without binarization: mining attribute profiles. PKDD 2006. *Lect. Notes Artif. Intell.*, **4213**, 528–535.

Hughes,T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

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

Ji,L and Tan,K.L. (2004) Mining gene expression data for positive and negative co-regulated gene clusters. *Bioinformatics*, **20**, 2711–2718.

Kanehisa,M. *et al.* (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.

Lee,H.K. *et al.* (2005) ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, **6**, 269.

Leung,Y.F. and Cavalieri,D. (2003) Fundamentals of cDNA microarray data analysis. *Trends Genet.*, **19**, 649–659.

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

Pasquier,N. *et al.* (1999) Discovering frequent closed itemsets for association rules. *Lecture Notes in Computer Science*, **1540**, 398–416.

Prelic,A. *et al.* (2005) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122–1129.

Quackenbush,J. (2001) Computational analysis of microarray data. *Nat. Genet.*, **2**, 418–427.

Raser,J.M. and O'Shea,E.K. (2005) Noise in gene expression: origins, consequences, and control. *Science*, **309**, 2010–2013.

Raychaudhuri,S. (2000) Principal components analisys to summarize microarray experiments: application to sporulation time series. Pac. Symp. Biocomput, 455–466.

Tamayo,P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoetic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Tavazoie,S. *et al.* (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.

Vohradsky,J. (2001) Neural network model of gene expression. *FASEB J.*, **15**, 846–854.

Zakharkin,S.O. *et al.* (2005) Sources of variation in Affymetrix microarray experiments. *BMC Bioinformatics*, **6**, 214.

Zaki,M.J. and Hsio,C.-J. (1999) *Charm: an efficient algorithm for closed itemset mining. Technical report*. Rensselaer Polytechnic Institute, Troy, NY.

Xu,X. *et al.* (2006) Mining shifting-and-scaling co-regulation patterns on gene expression profiles. *Intl. Confl. Data Eng.*, **00**, 89–98.