BRIEFINGS IN BIOINFORMATICS. VOL 11. NO 2. 210–224
Advance Access published on 8 October 2009

doi:10.1093/bib/bbp042

Gene association analysis: a survey of frequent pattern mining from gene expression data

Ronnie Alves, Domingo S. Rodriguez-Baena and Jesus S. Aguilar-Ruiz

Submitted: 27th April 2009; Received (in revised form): 5th August 2009

Abstract

Establishing an association between variables is always of interest in genomic studies. Generation of DNA microarray gene expression data introduces a variety of data analysis issues not encountered in traditional molecular biology or medicine. Frequent pattern mining (FPM) has been applied successfully in business and scientific data for discovering interesting association patterns, and is becoming a promising strategy in microarray gene expression analysis. We review the most relevant FPM strategies, as well as surrounding main issues when devising efficient and practical methods for gene association analysis (GAA). We observed that, so far, scalability achieved by efficient methods does not imply biological soundness of the discovered association patterns, and vice versa. Ideally, GAA should employ a balanced mining model taking into account best practices employed by methods reviewed in this survey. Integrative approaches, in which biological knowledge plays an important role within the mining process, are becoming more reliable.

Keywords: *gene expression analysis; gene association analysis; frequent pattern mining*

INTRODUCTION

It is widely believed that thousands of genes and their products (i.e. RNA and proteins) in a given living organism function in a complicated and orchestrated way. However, classical methods in molecular biology generally worked on a 'one gene in one experiment' basis and it implies a very limited throughput so the overall picture of gene function is hard to accomplish. In the past several years, a new technology, called DNA microarray, has attracted tremendous interests among biologists. The DNA microarray allows parallel genome-wide gene expression measurements of thousands of genes at a given time, under a given set of conditions and for cells/tissues of interest. Generation of microarray data introduces a variety of data analysis issues not encountered in traditional molecular biology or medicine. The data obtained from a series of microarray experiments is commonly in the form of an $N \times M$ matrix of expression levels, where the N rows correspond to various experimental conditions (generally hundreds) and the M columns correspond to genes under study (generally thousands).

Clustering and biclustering techniques are one of the most used computational strategies for analyzing microarrays [1, 2]. However, determining the interactions that can exist between different genes is not easily achieved by direct (bi)clustering solutions, particularly because genes can participate in more than one gene network. Thus, relationships that could be identified by gene association analysis (GAA) provide associations which do not appear adjacent to each other in a one shot clustering strategy [3–5].

Corresponding author. Ronnie Alves. CNRS UMR 6543, Institute of Developmental Biology and Cancer, Centre de Biochimie, Faculte des Sciences, 06108 Nice cedex 2. Tel: +33 4 92 07 69 47. E-mail: alves@unice.fr

Ronnie Alves is a postdoc at the Institute of Signaling Developmental Biology and Cancer at the University of Nice. He works on data mining methods for transcriptomics studies, with emphasis on the frequent pattern mining techniques.

Domingo Rodriguez-Baena is an assistant professor at Pablo de Olavide University, Seville, Spain. His main interests include data mining techniques, as biclustering and clustering, applied to gene expression datasets.

Jesus S. Aguilar-Ruiz is an associate professor at Pablo de Olavide University, Seville, Spain. He leads a research group on data mining and bioinformatics, and has published over 150 papers in international conferences and journals. Currently, he is the Dean of the School of Engineering and the co-Editor-in-Chief of the BioData Mining journal.

GAA is employed through the application of sophisticated association mining methods. These associations, usually represented in terms of implication rules, describe how the expression of one gene may be linked or associated with the expression of a set of genes. Besides, it is also possible to generate gene networks from discovered associations [6]. During the last decade the research community has focused on association mining methods since they not only reveal interesting gene relationships, but also are useful in integrative genomic studies [7]. In this sense, gene associations are also evaluated according to its linkage to other information obtained from several heterogeneous biological data sources.

Independently of applying GAA on a single source (microarray) or on an enriched one (microarray plus other biological information), finding interesting gene associations is not a trivial task. The intrinsic characteristics of the microarrays also bring the curse-of-dimensionality dilemma to GAA, and it is even more remarkable when one incorporates other biological information to enrich the final data model. In this survey, we focus on GAA from a frequent pattern mining (FPM) perspective. FPM is related to the most costly task of association mining methods, namely the enumeration of all possible combinations of gene pairs. Next, the extraction of rules from frequent subsets of genes is straightforward.

This work focuses on three aspects: strategies, scalability and biological soundness of the discovered patterns. These issues are essential to achieve efficient and practical GAA. Therefore, in this study we will survey main concepts and issues, data structures and algorithms that have already been proposed for exploring associations on DNA microarrays, as well as the most used strategies for evaluating biological soundness of discovered patterns.

This article is organized as follows: in ‘Association rules’ section, the main concepts related to association rules are described. FPM strategies are presented in ‘Mining frequent pattern’ section. ‘Using external biological information’ section deals with well-known biological data sources in order to improve and evaluate the quality of discovered gene association patterns, including examples of the associations extracted from data. An overall compilation of FPM methods and their strategies to extract biological knowledge is presented in ‘Summary of FPM methods for GAA’ section. Finally, in ‘Conclusions’ section the most interesting conclusions and future directions are summarized.

ASSOCIATION RULES

Concepts

Association rules have been extensively used with the aim of describing interesting relationships between variables in large datasets [8]. Next, the original association rule definition proposed by [9] is presented.

Definition 1 (Association Rule). *Let* $I = \{i_1, i_2, \ldots, i_m\}$ *be a set of* m *elements called* items. *A rule is defined as an implication of the form* $X \longrightarrow Y$, *where* $X, Y \subseteq I$ *and* $X \cap Y = \emptyset$. *The left-hand side of the rule is named* antecedent *and the right-hand side is named consequent.*

A typical application of association rules is the analysis of the so-called supermarket basket data, in which the goal is to find regularities in the customer behavior in terms of combinations of products that are purchased often together. In bioinformatics, association rules can be used to reveal biologically relevant associations between different genes, between environmental conditions and gene expression or even between biological information about genes and gene expressions [4]. For instance, an association rule between genes in the form $gene_A \longrightarrow gene_B, gene_C$ could mean that when gene A is overexpressed it is also very likely to observe an overexpression of genes B and C.

Hypothesis formulated on the validity of some concrete associations rules can be verified by correlation coefficients, which provide a numerical estimate of the association between two variables. Hence, these are used to assess the association between two gene expression profiles or to establish a connection between two genes in a genetic network [10]. *Pearson's Product Moment Correlation* (Pearson's rho), *Spearman's Rank-order Correlation Coefficient* (Spearman's rho) and *Kendall's Tau* are some of the most used correlation coefficients.

Although they can help biologists to test a concrete association between two variables, in studies based on microarrays there are huge datasets and little prior knowledge about possible relationships between variables. So, other methods, like FPM techniques, are very useful to explore over the intrinsic relations of data and to extract rules that provide a better understanding of genes behavior and their subsequent interactions.

FPM concepts

Frequent itemsets play an essential role in many data mining tasks. They are related to interesting patterns in datasets, such as association rules.

Definition 2 (Transaction) *Let $T = \{t_1, t_2, \ldots, t_n\}$ be a set of n subsets of items called transactions. Each transaction in T identifies a subset of items.*

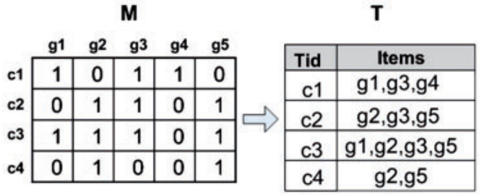
Definition 3 (Support of an itemset) *The support of an itemset X, support(X), is defined as the number of transactions in T which contain the itemset X.*

$$support(X) = \{t \in T | X \subseteq t\}|$$

Definition 4 (Frequent Itemset) *Given a set of items $I = \{i_1, i_2, \ldots, i_m\}$ and a set of transactions $T = \{t_1, t_2, \ldots, t_n\}$, a subset of I, $S \subseteq I$, is called a frequent itemset if S occurs in a percentage of all transactions in T that exceeds a threshold, named minimum support.*

FPM techniques provide methods to extract automatically all the frequent itemsets from a dataset in order to generate association rules from them. The problem of enumerating the number of maximal frequent itemsets in a dataset of transactions, given an arbitrary support threshold, is an extremely costly task [11]. As the search space contains exactly $2^{|I|}$ different itemsets, if I is large enough, then the naive approach to generate and count the supports of all itemsets over the dataset cannot be achieved within reasonable time. Therefore, the task of discovering all frequent itemsets is quite challenging.

In order to understand how to apply FPM algorithms in the context of GAA, FPM concepts will be related to gene expression data. What exactly constitutes an item or a transaction depends on the application and on the type of information to be extracted [4, 5]. Commonly, the meaning of transaction in terms of gene expression data is associated with ‘overexpression’, i.e. only those overexpressed genes will be understood to be included in the transaction. Nevertheless, other valid strategies could be ‘underexpression’, ‘downregulation’, ‘upregulation’ or involving time frames for further associations. Equivalently, the term frequent itemset is related to frequent subset of genes. Figure 1 clarifies the terminology, in which table T presents the items (genes) that take part of every transaction (experimental condition), only when they are overexpressed.



M

	g1	g2	g3	g4	g5
c1	1	0	1	1	0
c2	0	1	1	0	1
c3	1	1	1	0	1
c4	0	1	0	0	1

T

Tid	Items
c1	g1,g3,g4
c2	g2,g3,g5
c3	g1,g2,g3,g5
c4	g2,g5

Figure 1: *M* is the gene expression matrix, previously discretized (for instance, 1 means overexpressed, and 0 underexpressed). Experimental conditions *C* are in rows, whereas genes *G* are in columns. *T* is the set of transactions. Each transaction c_i contains a subset of items (genes).

Association rules extraction process

FPM tasks are only a part in the overall association rules extraction process. The general schema of this process is presented in Figure 2. The starting point is a $N \times M$ matrix of gene expression values, where the rows correspond to experimental conditions and columns represent genes. In *Phase 1*, the matrix is preprocessed. One reason for this data transformation deals with the adaptation of gene expression values to association rules mining, as these methods work with binary values (discretization). A key factor for determining which items belong to a certain transaction concerns with gene expression properties encoding [12]. Different expression properties might be considered, such as overexpression, up- or down-regulation or strong variation in order to determine the items of a certain transaction (Figure 1). For this purpose one can use statistical methods to detect differentially expressed genes, create different partitions by means a fixed threshold [5, 13], or apply adaptive discretization methods based on dynamic threshold selection policy [14].

Phase 2 has to do with the most costly task of association mining methods: FPM. The main aim is to extract all the frequent itemsets which *support* exceeds a certain threshold given by the users (see Definitions 3 and 4). The two most popular strategies are: column–enumeration-based methods (take the combination of genes as search space) and *row-enumeration-based methods* (search through the experimental conditions space). Due to the importance of these methods, they are discussed in ‘Mining frequent pattern’ section.

Once all the frequent itemsets have been obtained, the generation of association rules is performed in *Phase 3*. Any frequent itemset I of size greater than one is divided into two itemsets: X and Y, which will form the association rule $X \longrightarrow Y$ if its *support* exceeds a given threshold

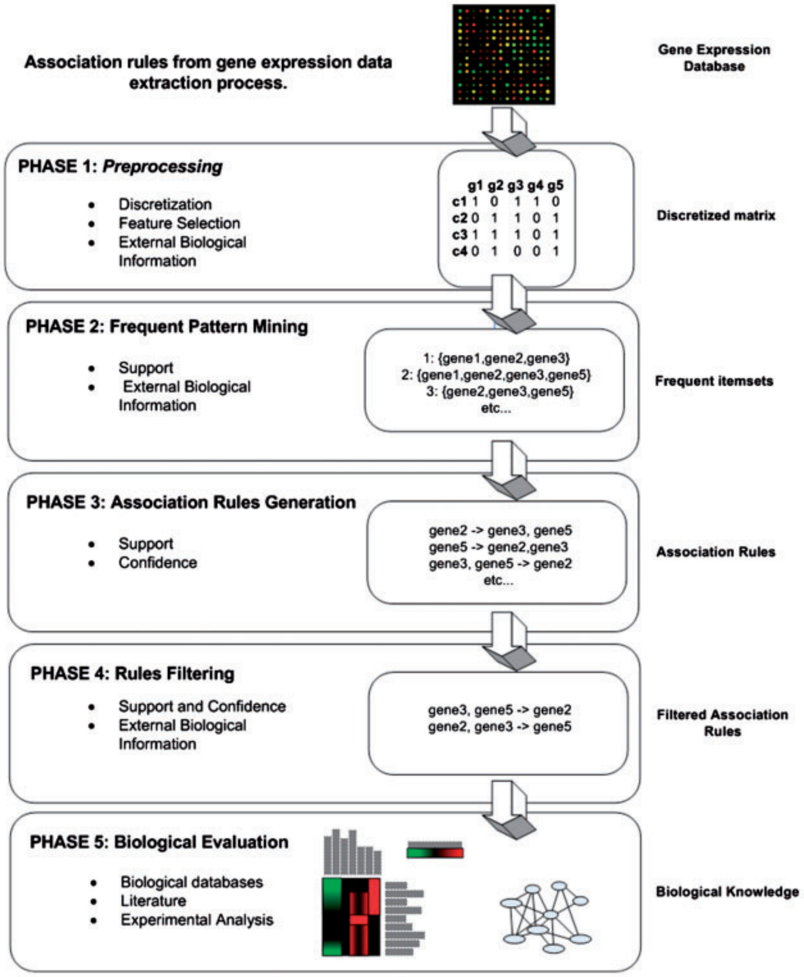


Figure 2: Association rules from gene expression data extraction process. It is composed of five sequential phases: initial dataset preprocessing, FPM techniques application in order to obtain frequent itemsets, generation of association rules from frequent itemsets, rules filtering and finally, biological evaluation to generate useful knowledge.

(Definition 5) and its *confidence* is high (Definition 6). According to the type of GAA application, those itemsets can represent relationships among different semantic concepts, like genes, annotations, other genomic information or even a labeled class (cancer, noncancer).

Definition 5 (Support of a rule) *The support of rule $X \longrightarrow Y$ with respect to the transaction set T is given by the ratio:*

$$support(X \longrightarrow Y) = \frac{support(X \cup Y)}{|\, T \,|}$$

Definition 6 (Confidence of a rule) *The confidence of rule $X \longrightarrow Y$ with respect to a transaction set T is given by the ratio:*

$$confidence(X \longrightarrow Y) = \frac{support(X \cup Y)}{support(X)}$$

Given a frequent itemset I, $2^k - 2$ association rules can be generated from it, being $k = |I|$. Hence, the *confidence* helps to reduce the number of association rules obtained, by selecting those with higher credibility among which a certain level of significance given by the *support* is shown.

To manage the very large number of discovered association patterns they have to be filtered, grouped and organized. Therefore, *Phase 4* is a necessary step in order to allow researchers to focus only on the most interesting association rules. External biological information might also be used for classification purposes [15].

Finally, in gene expression studies, the rules have to be evaluated for verifying their biological significance. For this purpose, prior biological knowledge from the literature or open access biological databases are usually taken into account in *Phase 5*. In some cases, association rules have been modeled as gene interaction networks at this final stage of the whole process [6, 16].

MINING FREQUENT PATTERNS

As it was mentioned in 'Association rule' section, the FPM phase is the most costly part from the computational point of view. In microarray data analysis, the specific gene expression dataset structure (thousands of genes against only hundreds of experimental conditions) increases the frequent itemsets mining process complexity. Due to this fact, developing efficient FPM techniques to be applied to genomic studies has been an important challenge during the last years. In this section, the most important FPM methods are reviewed from two angles: column-enumeration-based methods and row-enumeration-based methods. Given the relevance of the discretization as previous step to generating frequent itemsets, a brief description of the main discretization strategies is outlined next.

Discretization

Most works related to the application of association rule mining on gene expression profiles still rely on discretization tasks before applying any data mining technique. Although discretization may imply 'loss of information', it also alleviates the noise dilemma [4, 5]. It is not the aim of this survey to focus on discretization methods, so only a brief description of the main strategies will be presented next (the reader can refer to [12, 17] for more information).

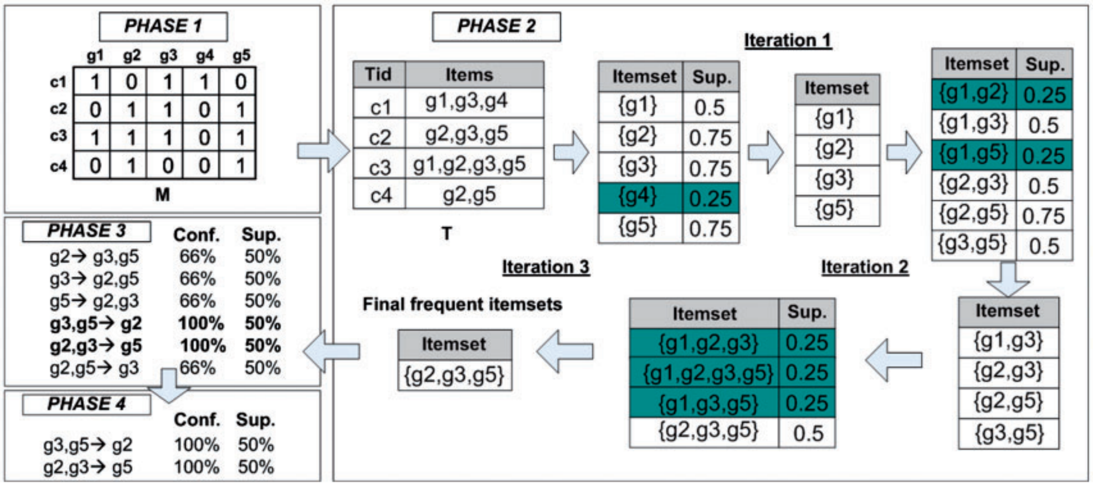
When the microarray dataset has a particular class associated to (for instance, tissue samples from cancer microarrays), the recursive minimization strategy proposed by Fayyad and Irani [18] is a suitable *supervised discretization* strategy. This method partitions the values of an attribute into a number of disjoint intervals in such a way that the entropy of the partition is minimal. Starting from a binary discretization boundary that minimizes the entropy function, a recursive algorithm is applied to both of the partitions created, until a stopping criteria is reached.

The basic unsupervised strategy is the *equal width partitioning*. It evolves sorting the observed values of a continuous feature and dividing the range of values for the variable into k equally sized bins, where k is a parameter supplied by the user. One can also make use of *equal depth partitioning*, in which each interval contains approximately the same number of values, or *equal width partitions*, as presented in [19]. Discretization over nonsupervised datasets also requires some prior evaluation concerning data distribution. Thus, it can be possible to find a suitable curve for binning such datasets. For this purpose one can use *threshold methods* [4, 5]. In this sense, genes with log expression values greater than a particular value are considered as overexpressed, otherwise as underexpressed. Recently, in [7] the authors proposed a *fuzzification* approach, partitioning the continuous domain into fuzzy sets. Such fuzzy-based model is likely to be more robust to noise when compared with other simple binning techniques. In [14], the thresholds are calculated dynamically by applying the same continuous-valued attribute discretization techniques as those used for classification algorithms based on decision trees.

In principle, the main issues about the use of discretization methods rely on two aspects: the distribution of data and the outliers processing. Once one can understand better the data distribution of a particular microarray, some assumptions can be taken in order to choose the more appropriate discretization method [20]. It is frequent to consider that data fits a normal distribution, although it is probably false. Some statistical tests (for instance, Pearson's chi-square test or Shapiro–Wilk test) are helpful to determine *normality* on data and then decide about which method is more suited. Regarding the detection of outliers, it is important to consider the impact of these on further discretization [21], hence smoothing their effects into the global data distribution will improve the quality of bins.

Column-enumeration-based strategy

Most of the proposed itemset-mining methods are a variant of the Apriori algorithm [9]. Apriori carries out a breadth-first search (BFS) that enumerates every single frequent itemset. Apriori also explores the downward closure property of an itemset filtering out non-frequent itemsets—the property that all subsets of a frequent itemset must be frequent. A simple Apriori example to provide a better understanding of FPM methods is illustrated in Figure 3. Let M be a discretized matrix, where 1 and 0 mean over- and under-expressed, respectively (Figure 1). Table T represents the transactions and their items. The Apriori algorithm performs a BFS through the search space of all the itemsets by iteratively generating candidate itemsets. At each iteration, the *support*



	g1	g2	g3	g4	g5
c1	1	0	1	1	0
c2	0	1	1	0	1
c3	1	1	1	0	1
c4	0	1	0	0	1

Tid	Items
c1	g1,g3,g4
c2	g2,g3,g5
c3	g1,g2,g3,g5
c4	g2,g5

Itemset	Sup.
{g1}	0.5
{g2}	0.75
{g3}	0.75
{g4}	0.25
{g5}	0.75

Itemset
{g1}
{g2}
{g3}
{g5}

Itemset	Sup.
{g1,g2}	0.25
{g1,g3}	0.5
{g1,g5}	0.25
{g2,g3}	0.5
{g2,g5}	0.75
{g3,g5}	0.5

Itemset
{g1,g3}
{g2,g3}
{g2,g5}
{g3,g5}

Itemset	Sup.
{g1,g2,g3}	0.25
{g1,g2,g3,g5}	0.25
{g1,g3,g5}	0.25
{g2,g3,g5}	0.5

Itemset
{g2,g3,g5}

PHASE 3	Conf.	Sup.
g2→ g3,g5	66%	50%
g3→ g2,g5	66%	50%
g5→ g2,g3	66%	50%
g3,g5→ g2	**100%**	**50%**
g2,g3→ g5	**100%**	**50%**
g2,g5→ g3	66%	50%

PHASE 4	Conf.	Sup.
g3,g5→ g2	100%	50%
g2,g3→ g5	100%	50%

Figure 3: Example of the Apriori algorithm, with support set to 2/4. Therefore, every itemset to be considered as a valid candidate must appear at least in two out of the four transactions. At each iteration, the support of candidate itemsets is calculated eliminating those which support is under the threshold. The removed itemsets are colored. The resulting itemsets are combined to create a new candidate group. At the end of the process, association rules are generated from the final frequent itemsets and their confidence and support are calculated. Finally, these rules are filtered following some criterion (e.g. confidence = 100%). In this example, Phase 1 (discretization), Phase 2 (FPM), Phase 3 (association rules generation) and Phase 4 (rule filtering) of the overall process described in previous section (Figure 2) are illustrated.

of every candidate itemset is calculated, eliminating those itemsets with *support* value under a threshold (set to 2/4 in this example). Based on the idea that an itemset is candidate if all its subsets are known to be frequent, the resulting itemsets are combined to create new candidate itemsets. The algorithm ends when no new candidate group can be generated. In Figure 3, after three iterations the final frequent itemset is composed of the genes g_2, g_3, g_5. From this frequent itemset the association rules are generated in phase 3. Afterwards in phase 4, the complete association rule set is filtered out following a given criterion (in Figure 3, the confidence is set to 100%, so only two out of six potential rules meet the filtering criterion).

Apriori-based methods show good performance with sparse datasets such as market–basket data, where the frequent patterns are very short. However, with dense datasets such as telecommunications, census data, microarrays, etc., where there are many long frequent patterns, these methods scale poorly and sometimes are impractical. This drawback is due to the high-computational cost of the evaluation of candidate and test sets used by Apriori-based approaches. Thus, new methods like Fp-Growth [22], which simplifies the problem of finding long patterns by concatenating small ones, have emerged as a promising strategy. In fact, several methods have been devised on the Fp-Growth basis [8, 23]. The main idea relies on a compact tree structure called Fp-tree, which is searched through recursively for enumerating all frequent patterns. The pattern growth is achieved by concatenating the suffix pattern with the frequent pattern generated from a conditional Fp-tree (for instance, the patterns with length equal to 1 will be used for generating those with length equal to 2, and so on). Even tree-based methods such as Fp-Growth may find some difficulties when dealing with high-dimensional datasets. A frequent pattern of size (number of items) s implies the presence of $2^s - 2$ additional frequent patterns as well, each of which is explicitly checked out by such methods. Thus, FPM algorithms that employ sophisticated heuristics for mining long frequent itemsets are practical solutions for GAA.

There are currently two alternatives for mining long patterns. The first one is to mine only maximal frequent itemsets, as in MaxMiner [24] and GenMax [25], which are typically orders of magnitude fewer than all frequent patterns. *Maximal itemsets* are those longest frequent patterns found under certain support threshold. Despite the fact that maximal patterns help understand the long itemsets in dense domains, they lead to loss of information; since subset counting is not available, maximal sets are not suitable for generating rules. The frequent set $\{g_2, g_3, g_5\}$ from Figure 3 is an example of a maximal frequent itemset. The second alternative is then mining only frequent closed sets as in CLOSE [26], CLOSET+ [27] and CHARM [28]. Closed sets are lossless in the sense that they can be used to uniquely determine the set of all frequent patterns and their exact frequencies. A 'closed itemset' is a frequent pattern that fits a support threshold and does not have any other super frequent pattern set with similar support value covering it (Figure 3). Examples of closed itemsets are the following ones: $\{g_3\}$, $\{g_1, g_3\}, \{g_2, g_5\}, \{g_2, g_3, g_5\}$. We can observe that the itemset $\{g_3, g_5\}$ is not a closed itemset since it is covered by the itemset $\{g_2, g_3, g_5\}$. Furthermore, closed-based algorithms can handle pattern redundancy, which is quite common in the application of association mining on high-dimensional databases [8, 23]. However, even by using such strategy the high dimensionality of microarrays still poses great challenges for column-enumeration-based methods.

Aforementioned methods employ exponential combination of all the columns (i.e. genes) in the gene expression matrix. Such search space size increases proportionally with the number of genes. Therefore, FPM methods that do not use candidate-set generation are usually more efficient. The type of patterns found also plays an important role in the strength or weakness of a FPM method. Thus, closed itemset strategies are more reliable for GAA. From such general discussion, one could expect that CLOSE+ is the most suitable column-enumeration approach for GAA. Indeed, the method was not applied to any kind of gene expression data, although it was successfully evaluated against its counterpart using other high dense datasets.

Row-enumeration-based strategy

Recently, support-based row-enumeration methods have emerged to handle efficiently GAA in microarrays. In terms of implementation it means that they use a vertical data format for enumerating frequent patterns rather than the horizontal format employed by most of the previously mentioned column-enumeration-based methods [23]. For instance in Figure 3, where each condition is a row and each gene is a column, the (row) enumeration process is then driven by intersecting the set of conditions instead of using the set of genes. As discussed previously, those classical column-enumeration methods might not be suitable for GAA, given the high dimensionality of DNA microarrays. Since the number of experiments (or experimental conditions) is lower than the number of genes in a microarray, new methods were proposed for enumerating frequent itemsets by considering the row-space (experiments) rather than the column-space (genes). These include CARPENTER [29], COBBLER [30], FARMER [31], TOPKRGS [32], TD-CLOSE [33], PATTERN-FUSSION [34] and MAXCONF [35].

CARPENTER [29] was the first method to explore the row-enumeration search space by constructing projected transposed tables recursively. Furthermore, it provides the complete frequent closed patterns. CARPENTER does recursive generation of conditional transposed tables, performing a depth-first traversal search of the row-enumeration tree. Pruning techniques, which were devised to enhance efficiency, prevent unnecessary traversal of the enumeration tree. A comparative study showed that CARPENTER improved CLOSET+ on about 500 times. Most of known row-enumeration algorithms have their basis on CARPENTER ideas.

FARMER [31] and TOPKRGS [32] were particularly designed to generate association rule classifiers of the form $X \longrightarrow C$, where C is a class label and X is a set of genes. These methods demand microarrays in which each experiment has a class associated to, e.g. 'cancer' and 'noncancer'. The itemset mining is supported through transposed tables taking into account that class information. Thus, each itemset in the transposed table should be enumerated accordingly to the positive and negative class. Both methods explore the idea of mining interesting groups of rules. For example, if we set a class label C to the row $= \{\text{c3: g1, g2, g3, g5} \longrightarrow \text{C}\}$, we could generate 15 rules in the form $X \longrightarrow C$, all of them covering the same row and having the same confidence (100%). Instead of generating all those rules, FARMER employs the concept of rule group, clustering them into a group with a unique upper bound ($\text{g1, g2, g3, g5} \longrightarrow \text{C}$, the most specific one) plus a set of lower bounds rules ($\text{g}[1:\text{n}]\, X \longrightarrow C$, the most general ones). Interestingness is reinforced by using user-specified thresholds like support, confidence and chi-square. TOPKRGS is quite similar to FARMER in terms of row-enumeration strategy of rule groups, but it differs in adopting a preference selection (top-k) to filter out significant rules, and its implementation using compact prefix-tree is more efficient.

Unlike Carpenter, Td-Close [33] develops a top-down row-enumeration method to search through the row-enumeration space, which makes the pruning power of minimum support threshold stronger than using bottom-up search style. Integrated with this search method, an effective and efficient closeness-checking strategy is also proposed.

The method called Maxconf [35] adopts an enumeration strategy similar to that of Carpenter, mining closed frequent sets and working through nonsupervised microarrays. It can extract interesting gene relationships with high confidence and low support. The previously mentioned row-enumeration methods cannot provide many interesting unknown gene relationships, since they rely entirely on the support measure to prune the search space. This is a great drawback as many potential gene associations, that have low support and high confidence, are filtered out by support-based methods. Maxconf may only fail at mining colossal patterns. Association mining tasks usually give greater importance to patterns that are longer in size. These large patterns, called *colossal patterns*, were first introduced in Pattern-Fusion [34].

Pattern-Fusion appears as a solution to deal with pattern explosion. It is based on the concept of core-pattern strategy, and also has an evaluation model proposed to assess the quality of the mining results against the complete set. Several studies conducted on both synthetic and real datasets demonstrated that Pattern-Fusion is able to give a good approximation for colossal patterns in datasets, unlike existing FPM algorithms. For instance, for the ALL-A dataset (cf. Table 1), this method discovered the largest colossal pattern with size greater than 85 genes, although the authors did not provide evidence of biological relevance.

Cobbler [30] is another FPM method that employs dynamic evaluation of closed itemsets by exploring either row-enumeration or feature-enumeration (column-enumeration) approach according to the dataset characteristics. Like Carpenter, it takes a depth-first traversal search from both trees with recursive construction of several conditional transposed tables. The switching condition is evaluated by estimating the enumeration costs for the subtrees and selecting the smallest one from both subtrees, i.e. feature or row based.

Unlike column-enumeration methods, the row-enumeration approaches prevent the itemset explosion by only expanding closed itemsets and enumerating the rows (experiments) instead of columns (genes). The row-enumeration space size is exponential with respect to the number of experiments. On the contrary, the column-enumeration space size is exponential with respect to the number of genes. Moreover, almost all the previous row-enumeration methods are support based, which means that low support might incur in combinatorial explosion, thus limiting the search for rare itemsets within the rule extraction phase. Maxconf addresses this issue with a free support-based strategy, composed of two levels of confidence pruning. A comparative analysis using several known datasets revealed that without using any support threshold Maxconf provided excellent results. The rules extracted, allowing two types of gene behavior (up and down), highlighted interesting relationships. For instance, the rule CSE1 → CRM1, PCL5, obtained from Hughes' dataset [36], with 100% confidence and 0.33% support, was biologically verified by the BIND database.

USING EXTERNAL BIOLOGICAL INFORMATION

Nowadays, researchers have a fast and easy access to biological information through the World Wide Web. For instance, PubMed Central (http://www.pubmedcentral.nih.gov/) is a digital archive of biomedical and life sciences journal literature created by the US National Institutes of Health (NIH). This scientific database contains research articles from more than 300 journals. The Gene Ontology (GO) [37] project provides a structured controlled vocabulary to describe gene and gene product attributes in any organism using three different types of ontologies (cellular component, molecular function and biological process). The Go database is a relational database comprising ontologies and annotations of genes and gene products related to terms in Go. The Kyoto Encyclopedia of Genes and Genomes (Kegg) [38] is a bioinformatics resource that integrates current knowledge on molecular interaction networks such as pathways and complexes as well as information about genes and proteins. The Biomolecular Interaction Network Database (Bind) [39] archives biomolecular interactions, reactions, complexes and pathway information.

The use of external information is a helpful strategy in any data mining task. Concretely, in association analysis of gene expression data prior biological knowledge can be used in many phases. Furthermore, biological information can be integrated in the gene expression database during the mining process. Finally, it is also important to verify if the final association rules generated are significant from a biological point of view. An association pattern is significant from a biological point of view when there is a significant set of genes in this rule that shares biological features. From biological databases, we extract the annotations related to those genes and measure, by means of statistical methods, the significant level of enrichment. Besides the biological evaluation, the most interesting associations might be true validated experimentally.

Next, the most representative examples of application of external biological information in GAA are presented. Although FPM can tackle different scenarios of GAA, in this work we focus on rule-based discovery as knowledge representation basis.

Quantitative association rules based on *half-spaces* are presented in [40], in which biological information is used as filtering basis for reducing database size. Such filtering is due to the fact that association rules based on *half-spaces* cannot afford the enumeration of all relevant rules, as the methods discussed in ‘Mining frequent patterns’ section, rather it uses an optimization approach. This optimization is carried out from observing rules through two hyperplanes (α and β). From a geometrical perspective, a hyperplane α is given by a vector $\overline{\alpha}$ and an intercept α_0. An instance x is then assigned to one half-space, if the dot product $\overline{\alpha}x + \alpha_0$ is positive and to the other half-space, if it is negative. In this work, the objective is to find rules defined by two hyperplanes: α, that specifies the condition of the left-hand side of the association rule, and β, that specifies the right-hand side. For example, one could build an association rule such as $ARG1 \times 0.99 - CAR1 \times 0.11 \geq 0.062 \rightarrow ARG3 > -0.032$, relating the expression levels of three genes in arginine metabolism. Only those rules with a high confidence score, that is, if it is located to the left of *alpha* and below *beta*, are susceptible to be generated.

In [4], the authors apply an integrative strategy to bring out interesting biological patterns. The method integrates biological knowledge and expression data, annotating genes with metabolic pathways from Kegg [38], transcriptional regulatory networks from literature and other annotations from the three categories of Go [37]. Thus, an association pattern is only reported when there is a significant set of genes that share biological features and similar expression patterns. In this sense, the associations are intrinsic to data, and further biological verification from other sources reinforces the potential significance of the associations. A similar strategy is proposed by [7], in which *fuzzy* rules are employed instead of the classical association rules. Fuzzy set theory is used to deal with microarray data, as it works well with imprecision and noise. Then, fuzzy rules strongly correlated with structural and functional gene features are extracted. Fuzzy association rules are expressions of the form $X \longrightarrow Y$, where X and Y are sets of fuzzy attribute-value pairs. For example, the authors are able to relate the genes GO annotations with their lengths generating rules like ‘GO = DNA helicase activity ⟶ length = HIGH’. An interesting example of fuzzy-based rule biclustering, containing 51 genes, is also highlighted in [7], in which gene associations are described as belonging to chromosome II and being annotated in the terms *macromolecules biosynthesis* and *cytosol*.

In [15], GAA is applied as a baseline for classification tasks. Thus, genes in rules are linked to a gene category during the mining process. These categories can be created according to various criteria (functionality, biochemical pathways, etc.) and they are useful for the filtering phase (Figure 2). Also, authors provide rule operators, including rule grouping, filtering, browsing and data inspection operators, to assist biologists on managing the very large number of discovered association patterns.

The use of biological knowledge in any of the phases of GAA enriches the final mining model and can help biologists to better understand genes and their complex relations. Using this information, many generated rules are confirmed to be known biological relationships among genes. For instance, the aforementioned Maxconf detects known direct biological interactions in Bind [39], and verifies if any of the gene interactions appears in specific chemical pathways. Sometimes, authors only rely on literature to find the suitable information for checking the biological soundness of their results [4, 5, 7, 40]. It is worth to mention that in biological validation, many generated rules should correspond to known biological relationships among genes. However, a noncorresponding rule does not imply incorrect relationship. It might be possible that this rule has not been hypothesized yet. Hence, these predictions should be biologically validated with new experimental analysis. A general view of classical measures of association patterns is given in [8, 23, 41].

The association rules extracted from gene expression data provide a very useful knowledge with different applications. For example, rules can be used as an additional support to the conclusions extracted from previous analysis. Thus, most of the rules extracted from gene expression [4, 5, 7, 35, 40] are confirmed by previous works. For instance, in [35] the rule: $\overline{MAC1}\longrightarrow\overline{FRE7}$, meaning that when gene *MAC1* is underexpressed gene *FRE7* is underexpressed too, was extracted from a yeast dataset [36]. The gene *MAC1* was selectively mutated in this dataset, and this rule correctly describes the relationship between the genes *MAC1* and *FRE7*. More specifically, *MAC1* activates the expression of the gene *FRE7* [42]. Therefore, *FRE7* cannot be expressed when *MAC1* is not, and this rule correctly indicates this causality. Moreover, rules bring the opportunity of formulating new hypothesis. In [5], using the information from different rules, authors discover that when the uncharacterized yeast genes NIT1 and YIL165C are expressed, then a very similar group of genes are expressed as well. Perhaps these two genes are biologically related, but this hypothesis is not confirmed by new experimental analysis.

Association rules can provide new biological knowledge beyond gene expression data relations, so interesting conclusions can be obtained by integrating information from different sources. For example, in [4] rules like: $Ribosome\longrightarrow [-]T6, [-]T7$ combines information about metabolic pathways, expression data and temporal dimension, meaning that genes involved in Ribosome pathway are underexpressed in time points 6 and 7. Authors in [7] take advantage of the recent availability of estimates of the protein amount and of the ability to change the expression level to extract rules like the following: $proteinAbundance = HIGH\longrightarrow G + C = HIGH$, that is, when the protein abundance is high, then the proportion of guanine plus cytosine in genes is high as well. There are rules that can show the temporal dependencies between the behavior of the genes. For example, rules generated in [43] represent various transcriptional time delays between associated genes: $POL30_{up}\,YLR183c_{up}\longrightarrow(14\ minutes)HTA2_{up}$ implies that the overexpression of genes POL30 and YLR183c from yeast is followed by the overexpression of HTA2 after 14 min. This approach, in which time frames are essential markers, searches for rules in the form X→(Δt)Y, where Δt is the temporal delay.

Detecting biological database errors is other application of association rules. In [3], authors were able to detect that the initial identification of a tag in SAGE (Serial Analysis of Gene Expression) data [44] was misleading. That is, the data mining technique presented in this article can allow the correct reassignment of wrongly data. Sometimes, FPM techniques are used for other purposes. GENECODIS [45] is a web-based tool to search for gene annotations and to rank them by statistical significance. In the process, APRIORI algorithm [9] is used to extract all combinations of annotations that appear in at least δ genes, with δ being a user-defined threshold. Thus, this tool provides not only gene annotations but also the potential relations among them.

SUMMARY OF FPM METHODS FOR GAA

In this section, all the FPM methods for GAA reviewed in this work are summarized. Table 1 provides an overall presentation of the set of methods as well as their strategies for enumerating highly correlated genes from gene expression data. This compilation is complemented with other characterization of the methods in terms of how biological information is processed during the knowledge discovery process (those methods with no application on GAA are not included in Table 1). For the sake of clarity, a chronological view of the FPM methods related to GAA discussed in this document is depicted in Figure 4.

A detailed evaluation with respect to either quality or efficiency of discovering biological patterns using any of the reviewed methods is outside the scope of this work, since it would require more complex controlled conditions. However, in general terms, we can assume that for association mining models in which candidates (potential gene pairs) are examined by implicitly or explicitly traversing a search tree in either depth-first or breadth-first fashion, if the search tree is exponential in size at some level, such exhaustive traversal search should run with exponential time complexity. Such

Table I: Overall compilation of FPM methods with direct application on GAA

Method	Patterns	Strategy	Dataset	Reference
Carpenter	Closed	Row	LC^1, ALL^2, OC^3	[29]
Cobbler	Closed	Row/Column	Synthetic, TH^4	[30]
Farmer	Rules	Row	$LC^1, BC^5, CT^6, PC^7, ALL - A^8$	[31]
TopKrgs	Rules	Row	$ALL - A^8, LC^1, OC^3, PC^7$	[32]
Half-Space	Rules	Half-Spaces	$SCE - H^9$	[40]
Td-Close	Closed	Row	LC^1	[33]
Integrative	Rules	Column	$SCEDS^{10}, HFSE^{11}$	[4]
Pattern-Fusion	All	Pattern-Fusion	$ALL - A^8$	[34]
Maxconf	Closed	Row	$SCE - H^9, SCE - M^{12}, SCE - S^{13}$	[35]
Fuzzy	Rules	Column	$SCE - ST^{14}$	[7]

#	Dataset name	Size	Access
1	Lung cancer	181 × 12000	http://www.chestsurg.org
2	Acute lymphoblastic leukemia	215 × 12000	http://www.stjuderesearch.org/data/ALLI
3	Ovarian cancer	253 × 15000	http://clinicalproteomics.steem.com
4	Compounds target to thrombin	1000 × 139000	http://www.biostat.wisc.edu
5	Breast cancer	97 × 24000	http://www.rii.com/publications
6	Colon tumor	62 × 2000	http://microarray.princetion.edu/oncology/affydata
7	Prostate cancer	136 × 12000	http://www-genome.wi.mit.edu/mpr/prostate
8	ALL-AML leukemia	38 × 886	http://www-genome.wi.mit.edu/cgi-bin/cancer
9	*Saccharomyces cerevisiae*	300 × 6000	[36]
10	SCE, diauxic shift (SCEDS)	63 × 6153	http://arep.med.harvard.edu/ExpressDB/arraydata/diauxic.newtxt
11	Human fibroblast after serum esposure	12 × 8000	http://genome-www.stanford.edu/serum
12	*Saccharomyces cerevisiae*	215 × 6000	[46]
13	*Saccharomyces cerevisiae*	82 × 6000	[47]
14	*Saccharomyces cerevisiae*	172 × 6152	http://www-genome.stanford.edu/yeast-stress/

The first column 'Method' presents the methods reviewed in this survey. The second and third columns are related to the strategy. Thus, 'Patterns' should be either the itemsets (All, Maximal or Closed ones) or rules. The enumeration strategy (Column or Row based) used to obtain such patterns is pointed on the third column. The microarray data characteristics used in the original papers are described in column 'Dataset' (which is detailed below). The notation used to define the size of every dataset is Rows × Columns, being rows the number of experimental conditions and columns the number of genes. The last column points out the reference of each method.

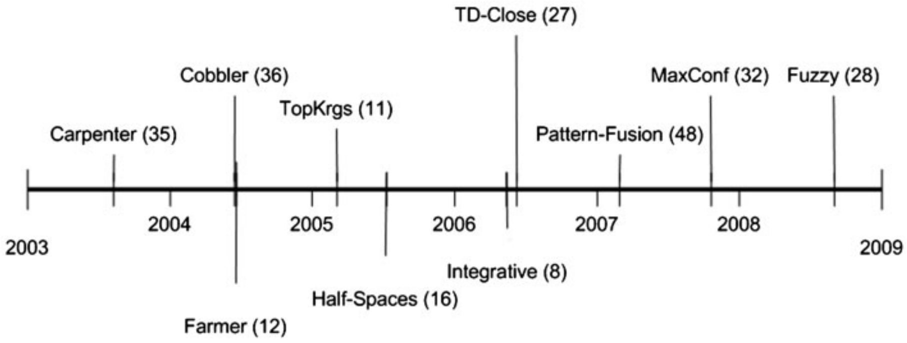


Figure 4: Chronogram of pattern frequent mining approaches.

observation is also the main motivation behind the Pattern-Fusion approach [34], which provides a good performance when the targets are long patterns.

Given the high dimensionality of microarrays, bioinformaticians have recently adopted row-enumeration methods as the most efficient strategy. Quality implies on having some biological evaluation before assuring biological soundness of found patterns. An interesting observation is that researchers are more concerned about scalability issues rather than quality. Most of the FPM methods, published in related peer-reviewed data mining conferences or journals, are quite robust on handling huge and sparse microarrays, while other methods published in bioinformatics circles, bring out more biological

Table 2: Biological knowledge used by FPM methods

Method	BioKnow	Application	AuxDB
Carpenter	None	Association	None
Cobbler	None	Association	None
Td-Close	None	Association	None
Pattern-Fusion	None	Association	None
Farmer	None	Classification	None
TopKrgs	After	Classification	Literature
Maxconf	After	Association	BIND, GO
Half-Space	Before/After	Association	Literature
Integrative	Before/After	Association	KEGG, GO
Fuzzy	Before/After	Association	GO, literature

knowledge from the discovered patterns. Ideally, new FPM methods for GAA should balance both scalability and validity in order to provide efficient and practical methods.

In general, all the methods reviewed are able to highlight interesting gene associations. However, the scenario of application can guide the choice following the type of pattern or the computational strategy. For example, TopKrgs and Farmer are recommended when the goal is to explore gene-to-target applications, where the target could be any conditional state in which genes are related to a particular biological study (cancer, noncancer). Other methods, like Carpenter, Cobbler, Td-Close, Maxconf and Pattern-Fusion are gene-to-gene applications, and are devised to find strong gene associations within the set of genes. Maxconf, with its free-support strategy, might provide higher confident patterns. On the other hand, strong and long association patterns are better addressed by Pattern-Fusion. In this sense, Half-Spaces is a good choice if some filtering step to alleviate the high dimensionality is introduced.

Table 2 shows how the reviewed FPM methods make use of biological knowledge to bring pattern interestingness. The first columns 'Method' stands for the FPM method, the second column 'BioKnow' describes how biological knowledge is integrated into the mining process (before or after mining) in order to enhance the discovered patterns. Some of these methods just take into account scalability rather than quality of the patterns (BioKnow = None). Only a few of the presented methods are built for classification purposes, but they also have their basis on association mining. This information is highlighted in the third column 'Application'. The last column 'AuxDB' refers to auxiliary information, which have been used to integrate, enhance or evaluate biological soundness of discovered patterns.

Table 3 shows different types of association rules extracted from gene expression data. As it can be observed, the table contains a wide variety of association rules. It is worth highlighting some of them. For example, the rule published in [40] is a quantitative rule of the form: *if the weighted sum of some variables is greater than a threshold then a second weighted sum of variables is greater than a second threshold.* In this case, the variables are the expression values of genes. Rule published in [7] is a fuzzy rule in which pairs of fuzzy attribute-value are used. Finally, rules published in [4] and [43] are related to temporal dependencies between gene expressions.

In short, the interest on integrating gene expression data with external biological databases in the context of GAA is growing, to the extent that the quality of results has turned into the main goal nowadays, leaving the efficiency issues in the background.

CONCLUSIONS

We have presented a survey on GAA in DNA microarray gene expression data. This work covers interesting issues related to GAA, from devising efficient computational strategies to the evaluation of biological pattern significance. The DNA microarray platforms generally provide highly correlated data. This observation impacts directly on how FPM methods should be designed in order to detect such correlations. So far, FPM methods based on row-enumeration strategies are those that can search efficiently for gene associations in microarrays. However, most FPM methods are not able to catch highly correlated unknown patterns, since they are

Table 3: Types of association rules extracted from gene expression data

Year	Reference	Rule	Rule description
2002	Becquet *et al*. [3]	Ribosomal 150 → Cytochrome 255	When gene encoding the *ribosomal protein S24* (identified by tag 150) is overexpressed, then gene encoding the *cytochrome c oxidase subunit IV* (identified by tag 255) is also overexpressed.
2003	Creighton *et al*. [5]	NIT1 → ATR1, BNA1, ...	When the gene NIT1 is overexpressed, then a group of genes are overexpressed as well.
2005	Georgii *et al*. [40]	−STE3 › 1.2 → −SAG1 › 1.1	Gene ST3 is underexpressed whenever SAG1 is underexpressed as well.
2006	Carmona *et al*. [4]	Ribosome → [-]T6, [-]T7	Genes involved in the metabolic pathway *Ribosome* are underexpressed in time points 6 and 7.
2007	McIntosh *et al*. [35]	$\overline{ESC8} \rightarrow \overline{IMD1}, \overline{IMD2}$	When gene ESC8 is underexpressed then genes IMD1 and IMD2 are underexpressed as well.
2008	Lopez *et al*. [7]	protein abundance = HIGH → G + C = HIGH	When the protein abundance is high, then the proportion of guanine plus cytosine in genes is high too.
2009	Nam *et al*. [43]	$POL30_{up}, YLR183C_{up} \rightarrow$ (14 minutes) $HTA2_{up}$	The overexpression of genes POL30 and YLR183c is followed by the overexpression of HTA2 after 14 min.

The first and second columns offer information referred to every specific research (year and bibliographic reference). The third column informs about the organism analyzed. The fourth column is an example of extracted rule and, finally, the last column is a briefly description of this rule.

mainly support-based approaches. Apart from having several FPM methods available for GAA, just a few of them really encompass biological knowledge. Many of the related FPM works concentrated their efforts on scalability issues and not in finding actionable biological patterns. On the other hand, new FPM methods have already been devised using the potential of related biomedical literature (PubMed) and scientific databases, such as GO database, KEGG and BIND, to discover interesting unknown biological knowledge. This is done either by employing an integrative approach or by pushing biological evaluation tasks either before or right after finding strong gene associations. There are also other types of gene expression data that are based on different platforms and biological processes, such as sequencing-based approaches like SAGE and Massively Parallel Signature Sequencing (MPSS), which will command more attention in the near future.

Key Points

- FPM is able to discover interesting association patterns in gene expression data.
- Several approaches have been designed for mining data, and promising applications are being developed or adapted for biological data.
- Due to the intrinsic properties of biological data, computational efficiency of FPM techniques is a key factor for discovering useful associations.
- The integration of biological knowledge into the mining process enriches the quality of discovered associations.
- Concepts and main strategies of FPM for gene expression analysis are reviewed.

Acknowledgements

We would like to thank reviewers for helpful suggestions and criticisms that have contributed to improve substantially the article.

FUNDING

Ministry of Science and Innovation and by Junta de Andalucia (Spain) (research grant number TIN2007-68084-C02-00); FCT (Fundação para Ciência e a Tecnologia) of Ministry for Science, Technology and Education of Portugal (grant SFRH/BD/17422/2004); Fundação Calouste Gulbenkian (FCG; grant 87749/2007).

References

1. Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: a Survey. *IEEE Trans Knowl Data Eng* 2004;**16**:1370–86.
2. Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput Biol Bioinform* 2004;**1**:24–45.
3. Becquet C, Blachon S, Jeudy B, *et al*. Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. *Genome Biol* 2002;**3**:12.
4. Carmona-Saez P, Chagoyen M, Rodriguez A, *et al*. Integrated analysis of gene expression by association rules discovery. *BMC Bioinformatics* 2006;**7**:54.
5. Creighton C, Hanash S. Mining gene expression databases for association rules. *Bioinformatics* 2003;**19**:79–86.
6. Bebek G, Yang J. PathFinder: mining signal transduction pathway segments from protein-protein interaction networks. *BMC Bioinformatics* 2007;**8**:335.
7. Lopez FJ, Blanco A, Garcia F, *et al*. Fuzzy association rules for biological data analysis: a case study on yeast. *BMC Bioinformatics* 2008;**9**:107.

8. Ceglar A, Roddick JF. Association mining. *ACM Comput Surv* 2006;**38**:2.

9. Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data,*. Washington, DC, USA: ACM Press, 1993;207–16.

10. Berrar DP, Dubitzky W, Granzow M, *et al*. *A Practical Approach to Microarray Data Analysis*, Vol. 3. Heidelberg, Germany: Springer 2006;12.

11. Yang G. The complexity of mining maximal frequent itemsets and maximal frequent patterns. In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. Seattle, Washington, USA: ACM Press, 2004;344–53.

12. Pensa RG, Leschi C, Besson J, *et al*. Assessment of discretization techniques for relevant pattern discovery from gene expression data. In: *Proceedings of 4th Workshop on Data Mining in Bioinformatics BIOKDD.* Seattle, Washington, USA, 2004;24–30.

13. Kotala P, Perera A, Zhou JK, *et al*. Gene expression profiling of DNA microarray data using peano count tree (p-trees). In: *Proceedings of the First Virtual Conference on Genomics and Bioinformatics*. North Dakota State University, USA, 2001; 15–16.

14. Ponzoni I, Azuaje F, Augusto JC, *et al*. Inferring adaptive regulation thresholds and association rules from gene expression data through combinatorial optimization learning. *IEEE/ACM Trans Comput Biol Bioinform* 2007;**4**: 624–33.

15. Tuzhilin A, Adomavicius G. Handling very large numbers of association rules in the analysis of microarray data. In: *Proceedings of the The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Edmonton, Alberta, Canada: ACM Press, 2002;396–404.

16. Hu M, Choi K, Su W, *et al*. A gene pattern mining algorithm using interchangeable gene sets for prokaryotes. *BMC Bioinformatics* 2008;**9**:124.

17. Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous features. In: *International Conference on Machine Learning*. Tahoe City, California, USA: Morgan Kaufmann, 1995;194–202.

18. Fayya and Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In: *Proceedings of the International Joint Conference on Uncertainty in AI*. Chambéry, France: Morgan Kaufmann, 1993;1022–27.

19. Zhang Z, Teo A, Chin-Ooi B, *et al*. Mining deterministic biclusters in gene expression data. In: *Proceedings of the 4th IEEE Symposium on Bioinformatics and Bioengineering*. Taichung, Taiwan: IEEE Computer Society, 2004;283–90.

20. Ismail M.K., Ciesielski V. An empirical investigation of the impact of discretization on common data distributions. In: *Proceedings of the Third International Conference on Hybrid Intelligent Systems*. Melbourne, Australia: IOS Press, 2003; 692–701.

21. Martinez R, Pasquier C, Pasquier N. GenMiner: mining informative association rules from genomic data. In: *Proceedings of the IEEE BIBM International Conference on Bioinformatics and Biomedecine*. Silicon Valley, CA, USA: IEEE Computer Society, 2007;15–22.

22. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. Dallas, Texas, USA: ACM Press, 2000;1–12.

23. Han J, Cheng H, Xin D, *et al*. Frequent pattern mining: current status and future directions. *Data Min Knowl Discov* 2007;**15**:55–86.

24. Bayardo RJ. Efficiently mining long patterns from databases. In: *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*. Seattle, Washington, USA: ACM Press, 1998;88–93.

25. Gouda K, Zaki MJ. GenMax: an efficient algorithm for mining maximal frequent itemsets. *Data Min Knowl Discov* 2005;**11**:223–42.

26. Pasquier N, Bastide Y, Taouil R, *et al*. Efficient mining of association rules using closed itemset lattices. *Inf Syst* 1999; **24**:25–46.

27. Wang J, Han J, Pei J. CLOSET+: searching for the best strategies for mining frequent closed itemsets. In: *Proceedings of the The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Washington, DC, USA: ACM Press, 2003;236–45.

28. Zaki MJ, Hsiao CJ. CHARM: an efficient algorithm for closed itemset mining. In: *Proceedings of the SIAM International Conference on Data Mining*. Arlington, VA, USA: SIAM Press, 2002;457–73.

29. Pan F, Cong G, Tung AK, *et al*. Carpenter: finding closed patterns in long biological datasets. In: *Proceedings of the the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Washington, DC, USA: ACM Press, 2003;637–42.

30. Pan F, Tung A, Cong G, *et al*. COBBLER: combining column and row enumeration for closed pattern discovery. In: *Proceedings of the 16th International Conference on Scientific and Statistical Database Management SSDBM*. Santorini Island, Greece: IEEE Computer Society, 2004; 21–30.

31. Cong G, Tung AK, Xu X, *et al*. FARMER: finding interesting rule groups in microarray datasets. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*. Paris, France: ACM Press, 2004; 143–54.

32. Cong G, Tan K, Tung AK, *et al*. Mining top-K covering rule groups for gene expression data. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*. Baltimore, Maryland, USA: ACM Press, 2005; 670–81.

33. Liu H, Han J, Xin D, *et al*. Top-down mining of interesting patterns from very high dimensional data. In: *Proceedings of the 22nd International Conference on Data Engineering*. Atlanta, GA, USA: IEEE Computer Society, 2006.

34. Zhu F, Yan X, Han J, *et al*. Mining colossal frequent patterns by core pattern fusion. In: *Proceedings of the IEEE 23rd International Conference on Data Engineering*. Istanbul, Turkey: ACM Press, 2007;706–15.

35. McIntosh T, Chawla S. High confidence rule mining for microarray analysis. *IEEE/ACM Trans Comput Biol Bioinform* 2007;**4**:611–23.

36. Hughes T, Marton M, Jones A, *et al*. Functional discovery via a compendium of expression profiles. *Cell* 2000;**102**: 109–26.

37. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. *Nat Genet* 2000;**25**:25–9.

38. Kanehisa M, Goto S, Hattori M, *et al*. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006;**32**:354–57.

39. Alfarano C, Andrade CE, Anthony K, Bahroos N, *et al*. The biomolecular interaction network database and related tools. *Nucleic Acids Res* 2005;**33**:D418–24.

40. Georgii E, Richter L, Rückert U. Analyzing microarray data using quantitative association rules. *Bioinformatics* 2005;**11**:123–9.

41. Tan P, Kumar V, Srivastava J. Selecting the right interestingness measure for association patterns. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Edmonton, Alberta, Canada: ACM Press, 2002;22–41.

42. Martins L, Jensen L, Simon J, *et al*. Metalloregulation of FRE1 and FRE2 homologs in Saccharomyces cerevisiae. *J Biol Chem* 1998;**273**:716–21.

43. Nam H, Lee K, Lee D. Identification of temporal association rules from time-series microarray data sets. *BMC Bioinformatics* 2009;**10**:3.

44. Wang SM. Understanding SAGE data. *Trends Genet* 2007; **23**:42–50.

45. Carmona-Saez P, Chagoyen M, Tirado F, *et al*. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol* 2007;**8**:R3.

46. Mnaimneh S, Davierwala AP, Haynes J, *et al*. Exploration of essential gene functions via titratable promoter alleles. *Cell* 2004;**118**:31–44.

47. Spellman P, Sherlock G, Zhang M, *et al*. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Biol Cell* 1998;**9**:3273–97.