Association Analysis Techniques for Bioinformatics Problems

Gowtham Atluri, Rohit Gupta, Gang Fang, Gaurav Pandey,
Michael Steinbach, and Vipin Kumar

Department of Computer Science and Engineering, University of Minnesota
{gowtham,rohit,gangfang,gaurav,steinbac,kumar}@cs.umn.edu
http://www.cs.umn.edu/~kumar/dmbio

Abstract. Association analysis is one of the most popular analysis paradigms in data mining. Despite the solid foundation of association analysis and its potential applications, this group of techniques is not as widely used as classification and clustering, especially in the domain of bioinformatics and computational biology. In this paper, we present different types of association patterns and discuss some of their applications in bioinformatics. We present a case study showing the usefulness of association analysis-based techniques for pre-processing protein interaction networks for the task of protein function prediction. Finally, we discuss some of the challenges that need to be addressed to make association analysis-based techniques more applicable for a number of interesting problems in bioinformatics.

Keywords: Data Mining, Association Analysis, Bioinformatics, Frequent Pattern Mining.

1 Introduction

The area of data mining known as association analysis[1] [1,2,50] seeks to find patterns that describe the relationships among the binary attributes (variables) used to characterize a set of objects. The iconic example of data sets analyzed by these techniques is market basket data, where the objects are transactions consisting of sets of items purchased by a customer, and the attributes are binary variables that indicate whether or not an item was purchased by a particular customer. The interesting patterns in these data sets are either sets of items that are frequently purchased together (frequent itemset patterns) or rules that capture the fact that the purchase of one set of items often implies the purchase of a second set of items (association rule patterns). Association patterns, whether rules or itemsets, are local patterns in that they hold only for a subset of transactions. The size of this set of supporting transactions, which is known as the support of the pattern, is one measure of the strength of a pattern. A key

[1] Not to be confused with the related, but separate field of statistical association analysis [3].

S. Rajasekaran (Ed.): BICoB 2009, LNBI 5462, pp. 1–13, 2009.

strength of association pattern mining is that the potentially exponential nature of the search can often be made tractable by using support based pruning of patterns [1], i.e., the elimination of patterns supported by too few transactions early on in the search process. Efforts to date have created a well-developed conceptual (theoretical) foundation [64] and an efficient set of algorithms [2,20]. The framework has been extended well beyond the original application to market basket data to encompass new applications [8,24,23,57].

Despite the solid foundations of association analysis and the potential economic and intellectual benefits of pattern discovery and its various applications, this group of techniques is not widely used as a data analysis tool in bioinformatics and computational biology. Some prominent examples of these data types are gene expression data [33] and data on genetic variations (e.g., single nucleotide polymorphism (SNP) data) [22]. Although the use of clustering and classification techniques is common for the analysis of these and other biological data sets, techniques from association analysis are rarely employed (The few exceptions include the work of researchers [5,13,30,29,40], including ourselves [57,37,35].). For instance, for the problem of protein function prediction, which is a key problem in bioinformatics [52], recent surveys [36,48,17] discuss several hundred papers using clustering and classification techniques, but only a handful using association analysis techniques. Thus, it has to be acknowledged that association analysis techniques have not found widespread use in this important domain.

In this paper we discuss some applications of association analysis techniques in bioinformatics and the challenges that need to be addressed to make these techniques applicable to other problems in this promising area. The rest of the paper is organized as follows: Section 2 presents a brief overview of various types of association patterns, which can be very useful for discovering different forms of knowledge from complex data sets, such as those generated by high-throughput biological studies. In the next section, we discuss a case study of how an association measure, $h - confidence$, can be used to address issues with the quality of the currently available protein interaction data. Section 3 discusses the use of association patterns for a bioinformatics application, namely addressing the noise and incompleteness issues with the currently available protein interaction network data. Section 4 provides concluding remarks and some of the challenges that needs to be addressed to extend the application of association patterns to a wide range of problems in bioinformatics.

2 Association Patterns

This section introduces some commonly used association patterns that have been proposed in the literature.

2.1 Traditional Frequent Patterns

Traditional frequent pattern analysis [50] focuses on binary transaction data, such as the data that results when customers purchase items in, for example, a grocery

store. Such market basket data can be represented as a collection of transactions, where each transaction corresponds to the items purchased by a specific customer. More formally, data sets of this type can be represented as a binary matrix, where there is one row for each transaction, one column for each item, and the ij^{th} entry is 1 if the i^{th} customer purchased the j^{th} item, and 0 otherwise.

Given such a binary matrix representation, a key task in association analysis is to finding frequent itemsets in this matrix, which are sets of items that frequently occur together in a transaction. The strength of an itemset is measured by its *support*, which is the number (or fraction) of transactions in the data set in which all items of the itemset appear together. Interestingly, support is an anti-monotonic measure in that the support of an itemset in a given data set can not be less than any of its supersets. This anti-monotonicity property allows the design of several efficient algorithms, such as Apriori [2] and FPGrowth [20], for discovering frequent itemsets in a given binary data matrix. However, an important factor in choosing the threshold for the minimum support of an itemset to be considered frequent is computational efficiency. Specifically, if n is the number of binary attributes in a transaction data set, there are potentially $2^n - 1$ possible non-empty itemsets. Since transaction data is typically sparse, i.e., contains mostly 0's, the number of frequent itemsets is far less than $2^n - 1$. However, the actual number depends greatly on the support threshold that is chosen. Nonetheless, with judicious choices for the support threshold, the number of patterns discovered from a data set can be made manageable. Also, note that, in addition to support, a number of additional measures have been proposed to determine the interestingness of association patterns [49].

2.2 Hyperclique Patterns

A hyperclique pattern [61] is a type of frequent pattern that contains items that are strongly associated with each other over the supporting transaction, and are quite sparse (mostly 0) over the rest of the transactions. As discussed above, in

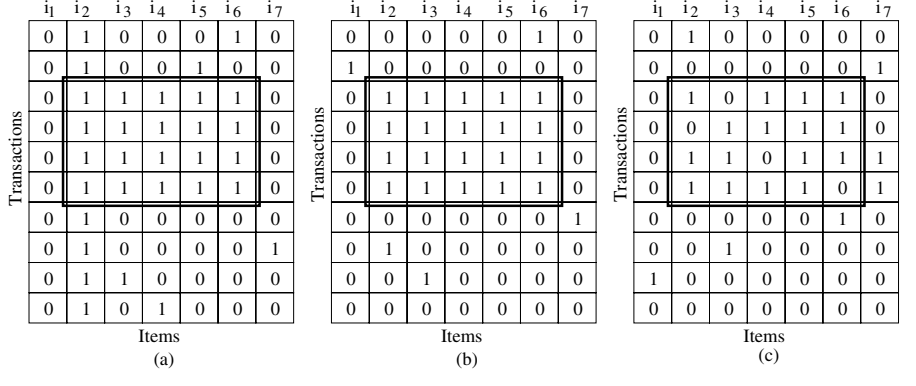


Fig. 1. Different types of association patterns (a) Traditional Frequent Patterns (b) Hyperclique Patterns (c) Error-tolerant Patterns

traditional frequent pattern mining, choosing the right support threshold can be quite tricky. If support threshold is too high, we may miss many interesting patterns involving low support items. If support is too low, it becomes difficult to mine all the frequent patterns because the number of extracted patterns increases substantially, many of which may relate a high-frequency item to a low-frequency item. Such patterns, which are called *cross-support patterns*, are likely to be spurious. For example, the pattern in Figure 1 (a), $\{i_2, i_3, i_4, i_5, i_6\}$ includes a high-frequency item i_2, which does not appear to have any specific association with other items in the patterns. Hyperclique patterns avoid these cross-support patterns by defining an anti-monotonic association measure known as *h-confidence* that ensures a high affinity between the itemsets constituting a hyperclique pattern [61]. Formally, the h-confidence of an itemset $X = \{i_1, i_2, \ldots i_m\}$, denoted as $hconf(X)$, is defined as,

$$hconf(X) = \frac{s(i_1, i_2, \ldots, i_k)}{max[s(i_1), s(i_2), \ldots, s(i_k)]}$$

where $s(X)$ is the support of an itemset X. Those itemsets X that satisfy $hconf(X) \geq \alpha$, where α is a user-defined threshold, are known as hyperclique patterns. These patterns have been shown to be useful for various applications, including clustering [60], semi-supervised classification [59], data cleaning [58], and finding functionally coherent sets of proteins [57].

2.3 Error-Tolerant Patterns

Traditional association patterns are obtained using a strict definition of support that requires every item in a frequent itemset to appear in each supporting transaction. In real-life datasets, this limits the recovery of frequent itemsets as they are fragmented due to random noise and other errors in the data. Motivated by such considerations, various methods [62,38,47,27,11] have been proposed recently to discover approximate frequent itemsets, which are also often called error-tolerant itemsets (ETIs). These methods tolerate some error by allowing itemsets in which a specified fraction of the items can be missing. This error tolerance can either be specified on the complete submatrix of the collection of items and transactions or in each row and/or column. For instance, in Figure 1(c), the itemset shown is a error tolerant itemset with 20% error tolerance in each row. It is important to note that each of the proposed definitions of error tolerant patterns will lead to a traditional frequent itemset if their error-tolerance is set to 0. For a detailed comparison of several algorithms proposed to discover ETIs from binary data sets, and their extensions, the reader is referred to our previous work [19].

2.4 Discriminative Pattern Mining

A variety of real-life data sets include information about which transactions belong to which of some pre-specified classes, i.e., class label information. For

such data sets, patterns of considerable interest are those that occur with disproportionate support or frequency in some classes versus the others. These patterns have been investigated under various names such as emerging patterns [16], contrast sets [4] and discriminative patterns [9,18,10], but we will refer to them as discriminative patterns. Consider the example in Figure 2, which displays a sample dataset, in which there are 14 items and two classes, each containing 10 instances (transactions). In this data set, four discriminative patterns can be observed: $P_1 = \{i_1, i_2, i_3\}$, $P_2 = \{i_5, i_6, i_7\}$, $P_3 = \{i_9, i_{10}\}$ and $P_4 = \{i_{12}, i_{13}, i_{14}\}$. Intuitively, P_1 and P_4 are interesting patterns that occur with differing frequencies in the two classes, while P_2 and P_3 are uninteresting patterns that have a relatively uniform occurrence across classes. Furthermore, we observe that P_4 is a discriminative pattern whose individual items are also highly discriminative, while P_1 is a discriminative pattern whose individual items are not.

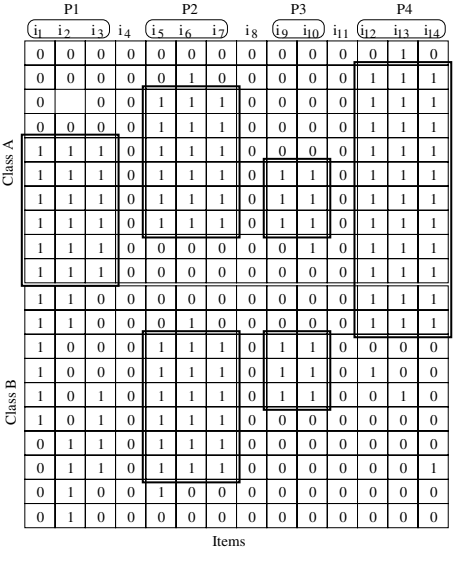


Fig. 2. Example of interesting discriminative patterns (P_1, P_4) and uninteresting patterns (P_2, P_3)

Discriminative patterns have been shown to be useful for improving the classification performance for transaction data sets when combinations of features have better discriminative power than individual features [9,55,53]. Discriminative pattern mining has the potential to discover groups of genes or SNPs that are individually not informative but are highly associated with a phenotype when considered as a group.

3 Case Study: Association Analysis-Based Pre-processing of Protein Interaction Networks

One of the most promising forms of biological data that are used to study the functions and other properties of proteins at a genomic scale are protein interaction networks. These networks provide a global view of the interactions between various proteins that are essential for the accomplishment of most protein functions. Due to the importance of the knowledge of these interactions, several high-throughput methods have been proposed for discovering them [25]. In fact, several standardized databases, such as DIP [56] and GRID [7] have now been

set up to provide systematic access to protein interaction data collected from a wide variety of experiments and sources.

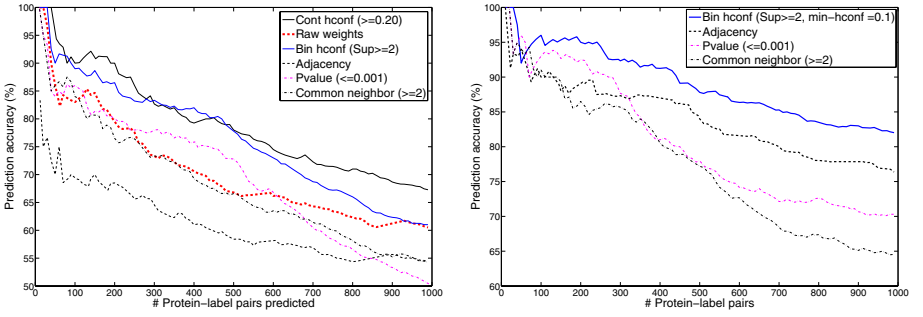
It is easy to see that a protein interaction network can be represented as an undirected graph, where proteins are represented by nodes and protein-protein interactions as edges. Due to this systematic representation, several computational approaches have been proposed for the prediction of protein function from these graphs [36,45,46,39,54,26,31]. These approaches can be broadly categorized into four types, namely neighborhoodbased, global optimization-based, clustering-based and association analysis-based. Due to the rich functional information in these networks, several of these approaches have produced very good results, particularly those that use the entire interaction graph simultaneously and use global optimization techniques to make predictions [31,54]. Indeed, recently, some studies have started using protein interaction networks as benchmarks for evaluating the functional relationships between two proteins, such as [63].

However, despite the advantages of protein interaction networks, they have several weaknesses which affect the quality of the results obtained from their analysis. The most prominent of these problems is that of noise in the data, which manifests itself primarily in the form of spurious or false positive edges [44,21]. Studies have shown that the presence of noise has significant adverse affects on the performance of protein function prediction algorithms [15]. Another important problem facing the use of these networks is their incompleteness, i.e., the absence of biologically valid interactions even from large sets of interactions [54,21]. This absence of interactions from the network prevents even the global optimization-based approaches from making effective use of the network beyond what is available, thus leading to a loss of potentially valid predictions.

A possible approach to address these problems is to transform the original interaction graph into a new weighted graph such that the weights assigned to the edges in the new graph more accurately indicate the reliability and strength of the corresponding interactions, and their utility for predicting protein function. The usefulness of hypercliques in noise removal from binary data [58], coupled with the representation of protein interaction graphs as a binary adjacency matrix to which association analysis techniques can be applied, motivated Pandey et al. [37] to address the graph transformation problem using an approach based on $h - confidence$ measure discussed earlier. This measure is used to estimate the common neighborhood similarity of two proteins P_1 and P_2 as

$$h - confidence(P_1, P_2) = \min\left(\frac{|N_{P_1} \cap N_{P_2}|}{|N_{P_1}|}, \frac{|N_{P_1} \cap N_{P_2}|}{|N_{P_2}|}\right) \quad (1)$$

where N_{P_1} and N_{P_2} denote the sets of neighbors of P_1 and P_2 respectively. As discussed earlier, this definition of $h - confidence$ is only applicable to binary data or, in the context of protein interaction graphs, unweighted graphs. However, the notion of h-confidence can be readily generalized to networks where the edges carry real-valued weights indicating their reliability. In this case, Equation 1 can be conveniently modified to calculate $h - confidence(P_1, P_2)$ by making the following substitutions: (1) $|N_{P_1}| \rightarrow$ sum of weights of edges incident on P_1 (similarly for P_2) and (2) $|N_{P_1} \cap N_{P_2}| \rightarrow$ sum of minimum of weights of each pair



(a) Results on the combined network (b) Results on the DIPCore network

Fig. 3. Comparison of performance of various transformed networks and the input networks (Best viewed in color and a larger size)

of edges that are incident on a protein P from both P_1 and P_2. In both these cases, the $h - confidence$ measure is guaranteed to fall in the range $[0, 1]$.

Now, with this definition, it is hypothesized that protein pairs having a high $h - confidence$ score are expected to have a valid interaction between them, since a high value of the score indicates a high common neighborhood similarity, which in turn reflects greater confidence in the network structure for that interaction. For the same reason, interactions between protein pairs having a low $h - confidence$ score are expected to noisy or spurious. Accordingly, Pandey *et al* [37] proposed the following graph transformation approach for pre-processing available interaction data sets. First, using the input interaction network $G = (V, E)$, the $h - confidence$ measure is computed between each pair of constituent proteins, whether connected or unconnected by an edge in the input network. Next, a threshold is applied to drop the protein pairs with a low $h - confidence$ to remove spurious interactions and control the density of the network. The resultant graph $G' = (V, E')$ is hypothesized to be the less noisy and more complete version of G, since it is expected to contain fewer noisy edges, some biologically viable edges that were not present in the original graph, and more accurate weights on the remaining edges.

In order to evaluate the efficacy of the resultant networks for protein function prediction, the original and the transformed graphs was provided as input to the FunctionalFlow algorithm [31], which is is a popular graph theory-based algorithm for predicting protein function from interaction networks. The performance was also compared with transformed versions generated using other common neighborhood similarity measures for such networks, such as Samanta et al [45]'s p-value measure. Figure 3 shows the performance of this algorithm on these transformed versions of two standard interaction networks, namely the *combined* data set constructed by combining several popular yeast interaction data sets (combined) and weighted using the EPR Index tool [14], and the other being a confident subset of the DIP database [14] (DIPCore). The performance is evaluated using the accuracy of the top scoring 1000 predictions of the functions of the constituent proteins generated by a five-fold cross-validation procedure,

where the functional annoitations are obtained from the classes at depth two of the FunCat functional hierarchy [43].

The results in Figure 3 show that for both the data sets, the $h - confidence$-based transformed version(s) substantially outperform the original network and the other measures for this task. The margin of improvement on the highly reliable DIPCore data set is almost consistently 5% or above, which is quite significant. Similar results are observed using the complete precision-recall curves. The interested reader is referred to [37] for more details on the methodology used and the complete results.

4 Concluding Remarks and Directions for Future Research

Association analysis has proved to be a powerful approach for analyzing traditional market basket data, and has even been found useful for some problems in bioinformatics in a few instances. However, there are a number of other important problems in bioinformatics, such as finding biomarkers using dense data like SNP data and real-valued data like gene-expression data, where such techniques could prove to be very useful, but cannot currently be easily and effectively applied.

An important example of patterns that are not effectively captured by the traditional association analysis framework and its current extensions, is a group of genes that are co-expressed together across a subset of conditions in a gene expression data set. Such patterns have often been referred to as *biclusters*. Figure 4 illustrates a classification of biclusters proposed by Madeira et al. [28]. They classified different types of biclusters into four categories: (i) biclusters with constant values (Figure 4(a)), (ii) biclusters with constant rows or columns (Example of a bicluster with constant rows is shown in Figure 4(b)), (iii) biclusters with coherent values, i.e., each row and column is obtained by addition or multiplication of the previous row and column by a constant value (Figure 4(c)), and (iv) biclusters with coherent evolutions, where the direction of change of values is important rather than the coherence of the values (Figure 4(d)). Each of these types of biclusters hold different types of significance for discovering important knowledge from gene expression data sets.

Since gene expression data is real-valued, traditional association techniques can not be directly applied since they are designed for binary data. Methods

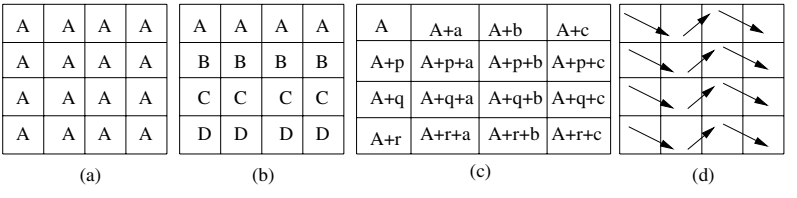


Fig. 4. Types of Biclusters: (a) Biclusters with constant values (b) Biclusters with constant rows (c) Biclusters with coherent values (additive model) (d) Biclusters with coherent evolutions

for transforming these data sets into binary form (for example, via discretization [5,13,30]) often suffer from loss of critical information about the actual. Hence, a variety of other techniques have been developed for and/or applied to this problem. These approaches include a wide variety of clustering techniques: ordinary partitional and hierarchical clustering, subspace clustering, biclustering/co-clustering, projective clustering, and correlation clustering. In addition, a variety of biclustering algorithms have been developed for finding such patterns from gene expression data, such as ISA [6], Cheng and Church's algorithm [12] and SAMBA [51], and more recently, for genetic interaction data [41]. Although these algorithms are often able to find useful patterns, they suffer from a number of limitations. The most important limitation is an inability to efficiently explore the entire search space for such patterns without resorting to heuristic approaches that compromise the completeness of the search. Pandey et al. [34] have presented one of the first methods for directly mining association patterns from real-valued data, particularly gene expression data, that does not suffer from the loss of information often faced by discretization and other data transformation approaches [34]. These techniques are able to discover all patterns satisfying the given constraints, unlike the biclustering algorithms that may only be able to discover a subset of these patterns. There are several open opportunities for designing better algorithms for addressing this problem.

Another challenge that has inhibited the use of association analysis in bioinformatics–even when the data is binary–is the density of several types of data sets. Algorithms for finding association patterns often break down when the data becomes dense because of the large number of patterns generated, unless a high support threshold is used. However, with a high threshold, many interesting, low-support patterns are missed. One particularly important category of applications with dense data are applications involving class labels, such as finding connections between genetic variations and disease. Consider the problem of finding connections between genetic variations and disease using binarized version of SNP-genotype data, which is 33% dense by design, since each subject must have one of the three variations of SNP pairs: *major-major*, *major-minor*, *minor-minor*. Traditional algorithms that do not utlize the class label information for pruning can only find patterns at high support, thus missing the low support patterns that are typically of great interest in this domain. In fact, most of the existing techniques for this problem only apply univariate analysis and rank individual SNPs using measures like p-value, odds ratio etc [3,22]. There are some approaches like Multi-Dimensionality Reduction (MDR) [42] and Combinatorial Partitioning Methods (CPM) [32], which are specially designed to identify groups of SNPs. However, due to their brute-force way of searching the exponential search space, these approaches also can only be applied to data sets with small number of SNPs (typically of the order of few dozen SNPs). Also, existing discriminative pattern mining algorithms [4,9,18,10] are only able to prune infrequent non-discriminative patterns, not the frequent non-discriminative patterns, which is the biggest challenge for dense data sets like SNP data and gene expression data. New approaches should be designed to

enable discriminative pattern mining on dense and high dimensional data, where effectively making use of class label information for pruning is crucial. Extension of association analysis based approaches to effectively use the available class label information for finding low-support discriminative patterns is a promising direction for future research.

In conclusion, significant scope exists for future research on designing novel association analysis techniques for complex biological data sets and their associated problems. Such techniques will significantly aid in realizing the potential of association analysis for discovering novel knowledge from these data sets and solve important bioinformatics problems.

References

1. Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases. In: Proc. SIGMOD, pp. 207–216 (1993)
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proc. VLDB, pp. 487–499 (1994)
3. Balding, D.: A tutorial on statistical methods for population association studies. Nature Reviews Genetics 7(10), 781 (2006)
4. Bay, S., Pazzani, M.: Detecting group differences: Mining contrast sets. DMKD 5(3), 213–246 (2001)
5. Becquet, C., et al.: Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human sage data. Genome Biology 3 (2002)
6. Bergmann, S., Ihmels, J., Barkai, N.: Iterative signature algorithm for the analysis of large-scale gene expression data. Physical Review 67 (2003)
7. Breitkreutz, B.-J., Stark, C., Tyers, M.: The GRID: the General Repository for Interaction Datasets. Genome Biology 4(3), R23 (2003)
8. Ceglar, A., Roddick, J.F.: Association mining. ACM Comput. Surv. 38(2), 5 (2006)
9. Cheng, H., Yan, X., Han, J., Hsu, C.-W.: Discriminative frequent pattern analysis for effective classification. In: Proc. IEEE ICDE, pp. 716–725 (2007)
10. Cheng, H., Yan, X., Han, J., Yu, P.: Direct mining of discriminative and essential graphical and itemset features via model-based search tree. In: Proc. ACM SIGKDD International Conference, pp. 230–238 (2008)
11. Cheng, H., Yu, P.S., Han, J.: Ac-close: Efficiently mining approximate closed itemsets by core pattern recovery. In: Proceedings of the 2006 IEEE International Conference on Data Mining, pp. 839–844 (2006)
12. Cheng, Y., Church, G.: Biclustering of Expression Data. In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology table of contents, pp. 93–103. AAAI Press, Menlo Park (2000)
13. Creighton, C., Hanash, S.: Mining gene expression databases for association rules. Bioinformatics 19(1), 79–86 (2003)
14. Deane, C.M., Salwinski, L., Xenarios, I., Eisenberg, D.: Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics 1(5), 349–356 (2002)
15. Deng, M., Sun, F., Chen, T.: Assessment of the reliability of protein–protein interactions and protein function prediction. In: Pac. Symp. Biocomputing, pp. 140–151 (2003)

16. Dong, G., Li, J.: Efficient mining of emerging paterns: Discovering trends and differences. In: Proceedings of the 2001 ACM SIGKDD International Conference, pp. 43–52 (1999)
17. Eisenberg, D., Marcotte, E.M., Xenarios, I., Yeates, T.O.: Protein function in the post-genomic era. Nature 405(6788), 823–826 (2000)
18. Fan, W., Zhang, K., Cheng, H., Gao, J., Yan, X., Han, J., Yu, P.S., Verscheure, O.: Direct discriminative pattern mining for effective classification. In: Proc. IEEE ICDE, pp. 169–178 (2008)
19. Gupta, R., Fang, G., Field, B., Steinbach, M., Kumar, V.: Quantitative evaluation of approximate frequent pattern mining algorithms. In: Proceeding of the 14th ACM SIGKDD Conference, pp. 301–309 (2008)
20. Han, J., Pei, J., Yin, Y., Mao, R.: Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery 8(1), 53–87 (2004)
21. Hart, G.T., Ramani, A.K., Marcotte, E.M.: How complete are current yeast and human protein-interaction networks? Genome. Biol. 7(11), 120 (2006)
22. Hirschhorn, J.: Genetic Approaches to Studying Common Diseases and Complex Traits. Pediatric Research 57(5 Part 2), 74R (2005)
23. Klemettinen, M., Mannila, H., Toivonen, H.: Rule Discovery in Telecommunication Alarm Data. J. Network and Systems Management 7(4), 395–423 (1999)
24. Kuramochi, M., Karypis, G.: An efficient algorithm for discovering frequent subgraphs. IEEE Trans. on Knowl. and Data Eng. 16(9), 1038–1051 (2004)
25. Legrain, P., Wojcik, J., Gauthier, J.-M.: Protein–protein interaction maps: a lead towards cellular functions. Trends Genet. 17(6), 346–352 (2001)
26. Lin, C., Jiang, D., Zhang, A.: Prediction of protein function using common-neighbors in protein-protein interaction networks. In: Proc. IEEE Symposium on BionInformatics and BioEngineering (BIBE), pp. 251–260 (2006)
27. Liu, J., Paulsen, S., Sun, X., Wang, W., Nobel, A., Prins, J.: Mining Approximate Frequent Itemsets In the Presence of Noise: Algorithm and Analysis. In: Proc. SIAM International Conference on Data Mining (2006)
28. Madeira, S.C., Oliveira, A.L.: Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans. Comput. Biol. Bioinf. 1(1), 24–45 (2004)
29. Martinez, R., Pasquier, N., Pasquier, C.: GenMiner: mining non-redundant association rules from integrated gene expression data and annotations. Bioinformatics 24(22), 2643–2644 (2008)
30. McIntosh, T., Chawla, S.: High confidence rule mining for microarray analysis. IEEE/ACM Trans. Comput. Biol. Bioinf. 4(4), 611–623 (2007)
31. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M.: Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics 21(suppl. 1), i1–i9 (2005)
32. Nelson, M., Kardia, S., Ferrell, R., Sing, C.: A Combinatorial Partitioning Method to Identify Multilocus Genotypic Partitions That Predict Quantitative Trait Variation. Genome Research 11(3), 458–470 (2001)
33. Nguyen, D.V., Arpat, A.B., Wang, N., Carroll, R.J.: DNA microarray experiments: biological and technological aspects. Biometrics 58(4), 701–717 (2002)
34. Pandey, G., Atluri, G., Steinbach, M., Kumar, V.: Association analysis for real-valued data: Definitions and application to microarray data. Technical Report 08-007, Department of Computer Science and Engineering, University of Minnesota (March 2008)

35. Pandey, G., Atluri, G., Steinbach, M., Kumar, V.: Association analysis techniques for discovering functional modules from microarray data. Nature Proceedings, Presented at ISMB, SIG Meeting on Automated Function Prediction (2008), http://dx.doi.org/10.1038/npre.2008.2184.1
36. Pandey, G., Kumar, V., Steinbach, M.: Computational approaches for protein function prediction: A survey. Technical Report 06-028, Department of Computer Science and Engineering, University of Minnesota (October 2006)
37. Pandey, G., Steinbach, M., Gupta, R., Garg, T., Kumar, V.: Association analysis-based transformations for protein interaction networks: a function prediction case study. In: Proceedings of the 13th ACM SIGKDD International Conference, pp. 540–549 (2007)
38. Pei, J., Tung, A., Han, J.: Fault-tolerant frequent pattern mining: Problems and challenges. In: Workshop on Research Issues in Data Mining and Knowledge Discovery (2001)
39. Pereira-Leal, J.B., Enright, A.J., Ouzounis, C.A.: Detection of functional modules from protein interaction networks. Proteins 54(1), 49–57 (2003)
40. Pfaltz, J., Taylor, C.: Closed set mining of biological data. In: Workshop on Data Mining in Bioinformatics (BIOKDD) (2002)
41. Pu, S., Ronen, K., Vlasblom, J., Greenblatt, J., Wodak, S.J.: Local coherence in genetic interaction patterns reveals prevalent functional versatility. Bioinformatics 24(20), 2376–2383 (2008)
42. Ritchie, M., et al.: Multifactordimensionality reduction reveals high-order iteractions among estrogen- metabolism genes in sporadic breast cancer. Am. J. Hum. Genet. 69(1), 1245–1250 (2001)
43. Ruepp, A., et al.: The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res. 32(18), 5539–5545 (2004)
44. Salwinski, L., Eisenberg, D.: Computational methods of analysis of protein-protein interactions. Curr. Opin. Struct. Biology 13(3), 377–382 (2003)
45. Samanta, M.P., Liang, S.: Predicting protein functions from redundancies in large-scale protein interaction networks. Proc. Natl. Acad Sci. U.S.A. 100(22), 12579–12583 (2003)
46. Schwikowski, B., Uetz, P., Fields, S.: A network of protein-protein interactions in yeast. Nature Biotechnology 18(12), 1257–1261 (2000)
47. Seppanen, J., Mannila, H.: Dense itemsets. In: KDD, pp. 683–688 (2004)
48. Seshasayee, A.S.N., Babu, M.M.: Contextual inference of protein function. In: Subramaniam, S. (ed.) Encyclopaedia of Genetics and Genomics and Proteomics and Bioinformatics. John Wiley and Sons, Chichester (2005)
49. Tan, P., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: Proceedings of the eighth ACM SIGKDD International Conference, pp. 32–41 (2002)
50. Tan, P.-N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison-Wesley, Reading (2005)
51. Tanay, A., Sharan, R., Shamir, R.: Discovering statistically significant biclusters in gene expression data. Bioinformatics 18(suppl. 1), 136–144 (2002)
52. Tramontano, A.: The Ten Most Wanted Solutions in Protein Bioinformatics. CRC Press, Boca Raton (2005)
53. van Vliet, M., Klijn, C., Wessels, L., Reinders, M.: Module-based outcome prediction using breast cancer compendia. PLoS ONE 2(10), 1047 (2007)

54. Vazquez, A., Flammini, A., Maritan, A., Vespignani, A.: Global protein function prediction from protein–protein interaction networks. Nat. Biotechnology 21(6), 697–700 (2003)
55. Wang, J., Karypis, G.: Harmony: Efficiently mining the best rules for classification. In: Proceedings of SIAM International Conference on Data Mining, pp. 205–216 (2005)
56. Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.-M., Eisenberg, D.: DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Research 30(1), 303–305 (2002)
57. Xiong, H., He, X., Ding, C., Zhang, Y., Kumar, V., Holbrook, S.R.: Identification of functional modules in protein complexes via hyperclique pattern discovery. In: Proc. Pacific Symposium on Biocomputing (PSB), pp. 221–232 (2005)
58. Xiong, H., Pandey, G., Steinbach, M., Kumar, V.: Enhancing data analysis with noise removal. IEEE Trans. on Knowl. and Data Eng. 18(3), 304–319 (2006)
59. Xiong, H., Steinbach, M., Kumar, V.: Privacy leakage in multi-relational databases via pattern based semi-supervised learning. In: Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 355–356. ACM, New York (2005)
60. Xiong, H., Steinbach, M., Tan, P., Kumar, V.: HICAP: Hierarchial Clustering with Pattern Preservation. In: Proceedings of the 4th SIAM International Conference on Data Mining, pp. 279–290 (2004)
61. Xiong, H., Tan, P.-N., Kumar, V.: Hyperclique pattern discovery. Data Min. Knowl. Discov. 13(2), 219–242 (2006)
62. Yang, C., Fayyad, U., Bradley, P.: Efficient discovery of error-tolerant frequent itemsets in high dimensions. In: Proc. ACM SIGKDD, pp. 194–203 (2001)
63. Yona, G., Dirks, W., Rahman, S., Lin, D.M.: Effective similarity measures for expression profiles. Bioinformatics 22(13), 1616–1622 (2006)
64. Zaki, M., Ogihara, M.: Theoretical foundations of association rules. In: 3rd ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (June 1998)