Association Analysis Techniques for Discovering Functional Modules from Microarray Data

Gaurav Pandey*, Gowtham Atluri, Michael Steinbach and Vipin Kumar
Department of Computer Science and Engineering, University of Minnesota, Twin Cities
*To whom correspondence should be addressed: gaurav@cs.umn.edu

1. INTRODUCTION

DNA microarray technology is one of the major recent advances in biotechnology [1]. In particular, their ability to measure the expression of thousands of genes simultaneously under a certain condition makes them suitable for several biological applications, such as the functional analysis of genes and identification of significantly over- or under-expressed genes in complex diseases like cancer. An application of great interest in microarray data analysis is the identification of a group of genes that show very similar patterns of expression in a data set, and are expected to represent groups of genes that perform common/similar functions, also known as functional modules [2]. Although clustering offers a natural solution to this problem, it suffers from the limitation that it uses all the conditions to compare two genes, whereas only a subset of them may be relevant.

Association analysis [3] offers an alternative route for finding such groups of genes that may be co-expressed only over a subset of the experimental conditions used to prepare the data set. The techniques in this field attempt to find groups of data objects that contain coherent values across a set of attributes, in an exhaustive and efficient manner. However, a major limitation of these techniques is that they are only able to analyze data sets that are constituted of binary and/or categorical variables, i.e., attributes whose values can come only from a finite set, such as {0,1}. Consequently, most of the applications of association analysis techniques to microarray data [4,5] incorporate a pre-processing step, wherein the expression values are discretized before groups of genes showing similar patterns of expression are extracted.

In recent work [6], we developed a generalization of association analysis for data sets where all the attributes are real-valued, such as gene expression datasets. More details of this methodology are provided in Section 2. This generalization enables us to efficiently extract groups of genes that show coherent patterns of expression across several experimental conditions, and are thus expected to constitute functional modules. We evaluate these modules for enrichment with a set of functional classes from the biological process ontology of GO. These evaluations show that a large number of the modules discovered are indeed functionally enriched, and thus demonstrate the ability of association analysis techniques to efficiently discover interesting functional information from microarray and other real-valued biological datasets. More details of these results are provided in Section 3.

2. METHODS

Several measures for the coherence of the values taken by a set of data objects across a set of cases have been defined in association analysis, mostly for binary data. One such measure is the *support* of a group of objects, also known as a *pattern* or an *itemset*, which measures the number of cases where each object in a group takes a value of 1. Also, much of the work on the design of efficient algorithms for extracting various types of patterns from data is based on the anti-monotonicity property of these association measures, particularly support. More specifically, the support of a group of objects I is guaranteed to be greater than or equal to that of a superset of this set. This property enables the design of algorithms such as Apriori [10], that are able to discover hundreds of such groups of coherent patterns that have a certain minimum support from large datasets very efficiently.

In order to generalize such analysis for large real-valued datasets, it is necessary to define a support measure that captures the coherence of the values of the data objects and is anti-monotonic. In particular, since our focus is on the use of these techniques for analyzing gene expression data, we took into account the following two aspects of the expression values to define such a support measure:

a) ***Direction***: Since we use the log_2-transformed ratio as the expression value of a gene, an important consideration for measuring the coherence of values of a group of genes in a given condition is the

sign or parity of the value. Thus, we only allow conditions under which all the genes in a pattern are either over- or under-expressed to contribute towards the coherence of the pattern.

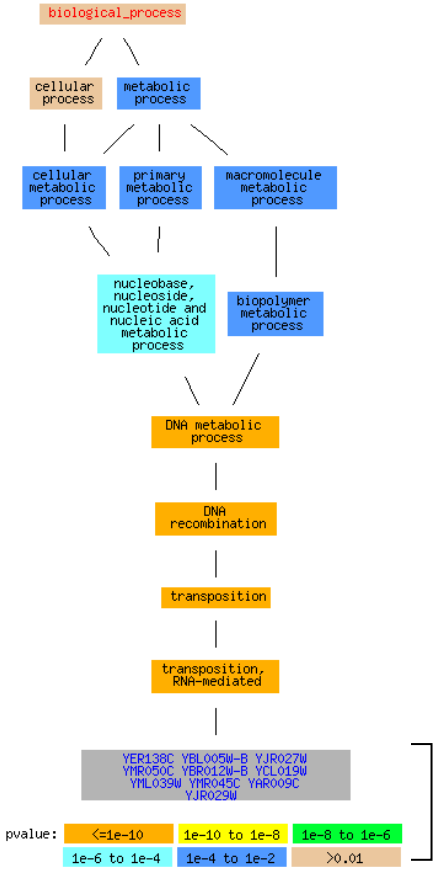
b) ***Magnitude***: Another important consideration for measuring the coherence of the expression of a group of genes is that the magnitude of their expression should be within a certain range. We incorporate this requirement in an adaptive manner, where this range is computed as a multiple of the expression value of the least expressed gene in the group under each condition.

Using these two semantic requirements for measuring the coherence of the expression values of a set of genes across several experimental conditions, we defined an anti-monotonic support measure named *ContSupport*, the mathematical details of which can be found in [6]. This measure also enabled the design of an efficient algorithm for deriving groups of genes from a microarray dataset that are highly co-expressed across a significant subset of conditions. We would like to note here that other methods of measuring the coherence of expression values of a set of genes are possible, and our definitions above represent one such method. In the next section, we discuss the evaluation of the functional enrichment of the set of gene patterns generated from a standard gene expression dataset.

3. RESULTS

We applied our gene pattern discovery algorithm to Hughes *et al* [7]'s popular *S. cerevisiae* microarray data set. The dimensions of this dataset are 300 conditions × 6316 genes, and thus, its large scale nature justifies the use of sophisticated data mining algorithms for extracting useful knowledge about functions of the constituent genes. The gene patterns extracted using various thresholds were tested for enrichment by Myers *et al*'s list [8] of 138 functional classes from the GO biological process ontology, and significantly enriched patterns were identified after correcting for multiple hypothesis testing using Bonferroni correction.

Figure 1 shows the annotation of a pattern of ten genes discovered by the *RNA-mediated transposition* class in the GO biological process ontology. The accurate annotation of this large set by a very specific functional class illustrates the ability of association analysis techniques to derive reasonably large groups of inter-related genes. The evaluation of a large set of patterns generated by this technique at different parameter values also produces similar positive results. For instance, among a set of 26093 patterns derived using certain thresholds, 8347 (31.98%) patterns were found to be significantly enriched by at least one functional class in the set considered, after Bonferroni correction. Also, the 500 largest patterns, which are usually of most interest to practitioners, are all found to be enriched by at least one of these classes with a *p-value* smaller than 10^{-10}. More detailed results can be found in [6].



In summary, these results illustrate that association analysis is a useful technique for deriving functional modules and other types of gene groups from large microarray gene expression datasets.

4. REFERENCES

1. D. J. Duggan et al. Expression profiling using cDNA microarrays. *Nature Genetics*, 21(1 Suppl):10–14, 1999.
2. M. B. Eisen et al. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci*, 95(25):14863-14868, 1998.
3. P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson Addison-Wesley, 2006.
4. C. Becquet et al. Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. *Genome Biology*, 3, 2002.
5. T. McIntosh and S. Chawla. High confidence rule mining for microarray analysis. *IEEE/ACM TCBB*, 4(4):611–623, 2007.
6. G. Pandey et al. Association Analysis for Real-valued Data: Definitions and Application to Microarray Data, TR 08-007, Department of Computer Science and Engineering, University of Minnesota, Twin Cities, submitted to KDD 2008.
7. T. R. Hughes et al. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, 2000.
8. C. L. Myers et al. Finding function: evaluation methods for functional genomic data. *BMC Genomics*, 7:187, 2006.
9. E. I. Boyle et al. Go::termfinder—open source software foraccessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, 2004.
10. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. VLDB*, pages 487–499, 1994.