



Mining gene expression databases for association rules

Chad Creighton^{1,*} and Samir Hanash²

¹Bioinformatics Program and ²Pediatrics and Communicable Diseases, University of Michigan, Ann Arbor, MI 48109, USA

Received on April 19, 2002; revised on July 1, 2002; accepted on July 10, 2002

ABSTRACT

Motivation: Global gene expression profiling, both at the transcript level and at the protein level, can be a valuable tool in the understanding of genes, biological networks, and cellular states. As larger and larger gene expression data sets become available, data mining techniques can be applied to identify patterns of interest in the data. Association rules, used widely in the area of market basket analysis, can be applied to the analysis of expression data as well. Association rules can reveal biologically relevant associations between different genes or between environmental effects and gene expression. An association rule has the form $LHS \Rightarrow RHS$, where LHS and RHS are disjoint sets of items, the RHS set being likely to occur whenever the LHS set occurs. Items in gene expression data can include genes that are highly expressed or repressed, as well as relevant facts describing the cellular environment of the genes (e.g. the diagnosis of a tumor sample from which a profile was obtained).

Results: We demonstrate an algorithm for efficiently mining association rules from gene expression data, using the data set from Hughes *et al.* (*Cell*, **102**, 109–126, 2000) of 300 expression profiles for yeast. Using the algorithm, we find numerous rules in the data. A cursory analysis of some of these rules reveals numerous associations between certain genes, many of which make sense biologically, others suggesting new hypotheses that may warrant further investigation. In a data set derived from the yeast data set, but with the expression values for each transcript randomly shifted with respect to the experiments, no rules were found, indicating that most all of the rules mined from the actual data set are not likely to have occurred by chance.

Availability: An implementation of the algorithm using Microsoft SQL Server with Access 2000 is available at <http://dot.ped.med.umich.edu:2000/pub/assoc.rules/assoc.rules.zip>. Our results from mining the yeast data set are available at <http://dot.ped.med.umich.edu:2000/pub/assoc.rules/yeast.results.zip>.

Contact: ccreight@umich.edu.

1 INTRODUCTION

Gene expression data, both at the transcript level and at the protein level, can be a valuable tool in the understanding of genes, biological networks, and cellular states. One goal in analyzing expression data is to try to determine how the expression of any particular gene might affect the expression of other genes; the genes involved in this case could belong to the same gene network. By a gene network, we mean a set of genes being expressed together in a non-random pattern. Another goal of expression data analysis is to try to determine what genes are expressed as a result of certain cellular conditions, e.g. what genes are expressed in diseased cells that are not expressed in healthy cells. While early experiments using microarrays profiled only a few samples, more recent experiments profile on the order of dozens or even hundreds of samples, allowing for a more robust statistical analysis of the data. In the near future, data sets containing thousands of samples should become available. As gene expression data sets become larger and larger, spreadsheets will become less and less of an adequate tool for doing analysis (as a single worksheet in Excel can hold no more than 256 columns), and data mining techniques using large databases should find more and more use in analyzing expression data.

Many clustering techniques for grouping genes based on similar expression profiles have been explored (Eisen *et al.*, 1998; Tavazoie *et al.*, 1999; Tamayo *et al.*, 1999). One common data mining technique, different from clustering, for finding and describing relationships between different items in a large data set is to look for *association rules* in the data. An association rule has the form $LHS \Rightarrow RHS$, where LHS and RHS are sets of items, the RHS set being likely to occur whenever the LHS set occurs. Association rules are used widely in the retail industry under the name ‘market basket analysis’. Association rules have been used as well to mine medical record data (Doddi *et al.*, 2001; Stilou *et al.*, 2001). In market basket analysis, an association rule represents a set of items

*To whom correspondence should be addressed.

that are likely to be purchased together; for example, the rule $\{cereal\} \Rightarrow \{milk, juice\}$ would state that whenever a customer purchases cereal, he or she is likely to purchase both milk and juice as well in the same transaction. In the analysis of gene expression data, the items in an association rule can represent genes that are strongly expressed or repressed, as well as relevant facts describing the cellular environment of the genes (e.g. a diagnosis for a tumor sample that was profiled, or a drug treatment given to cells in the sample before profiling). An example of an association rule mined from expression data might be $\{cancer\} \Rightarrow \{gene\ A \uparrow, gene\ B \downarrow, gene\ C \uparrow\}$, meaning that, for the data set that was mined, in most profile experiments where the cells used were cancerous, gene *A* was measured as being up (i.e. highly expressed), gene *B* was down (i.e. highly repressed), and gene *C* was up, altogether.

Public gene expression data sets large enough to mine for association rules and obtain meaningful results are already available. Algorithms for finding rules efficiently have been extensively developed in market basket analysis, and we apply a version of one of these algorithms to mine the compendium of Hughes *et al.* (2000) of profiles from 300 diverse mutations and chemical treatments in yeast. We find numerous rules in the data, a cursory analysis of some of which reveals numerous associations between certain genes, many of which make sense biologically, others suggesting new hypotheses that may warrant further investigation. In a data set derived from the yeast data set, but with the expression values for each transcript randomly shifted with respect to the experiments, no rules were found, indicating that very few of the rules mined from the actual data set are likely to have existed in the data by chance.

The remainder of this paper is organized as follows. In Section 2, we give a brief review of association rules, extending the concept as it could apply to gene expression data, and then describe an efficient algorithm for finding rules. In Section 3, we describe a database application (freely available, see **Abstract**) that we wrote to implement an algorithm for mining association rules from gene expression data. In Section 4 we describe the results of mining an expression data set for yeast. Conclusions and ideas for future applications of this method are provided in Section 5.

2 METHODS

2.1 Association rules described

An association rule has the form $LHS \Rightarrow RHS$, where *LHS* and *RHS* are *itemsets*. Itemsets can be defined in terms of *transactions*, which in the retail industry refer to customer transactions (a customer purchases one or more items at the checkout counter in a single transaction). Here we use

the following definition for itemsets and association rules (as provided by Doddi *et al.*, 2001):

DEFINITION 1.

1. Given a set *S* of items, any nonempty subset of *S* is called an ‘itemset’.
2. Given an itemset *I* and a set *T* of transactions, the ‘support’ of *I* with respect to *T*, denoted by $\text{support}_T(I)$, is the number of transactions in *T* that contain all the items in *I*.
3. Given an itemset *I*, a set *T* of transactions and a positive integer α , *I* is a ‘frequent itemset’ with respect to *T* and α if $\text{support}_T(I) \geq \alpha$. We refer to α as the ‘minimum support’.

DEFINITION 2.

1. An ‘association rule’ is a pair of disjoint itemsets. If *LHS* and *RHS* denote the two disjoint itemsets, the association rule is written as $LHS \Rightarrow RHS$.
2. The ‘support’ of the association rule $LHS \Rightarrow RHS$ with respect to a transaction set *T* is the support of the itemset $LHS \cup RHS$ with respect to *T*.
3. The ‘confidence’ of the rule $LHS \Rightarrow RHS$ with respect to a transaction set *T* is the ratio $\text{support}(LHS \cup RHS) / \text{support}(LHS)$.

In market basket analysis, frequent itemsets represent things that customers will often buy together, such as cereal and milk, denoted as $\{cereal, milk\}$. A set of items can be considered frequent if they occur in a percentage of all transactions that exceeds the minimum support criteria. From these frequent itemsets, we can derive rules such as $\{cereal\} \Rightarrow \{milk\}$, meaning that if a customer buys cereal, he or she is likely to buy milk in the same transaction. For the rule $\{cereal\} \Rightarrow \{milk\}$ to be derived from the frequent itemset $\{cereal, milk\}$, the rule should have a high confidence with respect to the data set, i.e. milk would need to have been purchased in a high percentage of the transactions in which cereal was purchased.

2.2 Association rules applied to gene expression data

In the context of market basket analysis, a gene expression profile can be thought of as a single transaction, and each transcript or protein can be thought of as an item. However, while in market basket analysis any particular item is either purchased or not purchased in a transaction, in an expression profile each transcript or protein is assigned a real value that specifies the relative abundance of that transcript or protein in the profiled sample. In applying association rules to gene expression data, one technique would be to first bin each measured value

as being up (i.e. highly expressed), down (i.e. highly repressed), or neither up nor down. In trying to determine the interactions between genes using expression profiles, one must account for a good deal of noise in the data, arising not only from measurement error, but from noise that is probably inherent to biological systems in general (Thattai and van Oudenaarden, 2001; Hughes *et al.*, 2000). We may not always expect slight fluctuations in the expression levels of one gene in a gene network to have tightly-coupled effects on the other genes. Binning the values is therefore one way to help alleviate problems with noise, allowing us to focus on the more general up/down effects of genes in a network. In this case, any particular gene in a profile can be thought of as being two ‘items,’ one item referring to the gene being up, the other referring to the gene being down. A gene expression profile ‘transaction’ would include the set of genes that were up and the set of genes that were down in the profile.

As well as the up and down states of genes, items in a gene expression profile transaction can include relevant facts describing the cellular environment. For example, the rule $\{heat\ shock\} \Rightarrow \{gene\ A\uparrow, gene\ B\downarrow\}$ could indicate that both gene *A* is up and gene *B* is down in most cases where a heat shock treatment is first given to the cells before profiling. In order to include a sample attribute in an expression data set for mining association rules, the attribute value could be binned in such a way as to be ‘up,’ ‘down,’ or neither up nor down. For example, for an *age* attribute that gives the age of a patient from which the sample was obtained, the attribute could be represented in that data set as an item called *age* > 60, which could have a value of ‘up’ if the patient’s age was over 60 and ‘down’ if otherwise.

2.3 Finding association rules

The first step in finding association rules is to look for frequent itemsets. A commonly used algorithm for doing this is the *Apriori* algorithm (Ramakrishnan and Gehrke, 2000). The algorithm relies upon a simple yet fundamental property of frequent itemsets, called the *a priori* property: Every subset of a frequent itemset must also be a frequent itemset. The algorithm proceeds iteratively, first identifying frequent itemsets containing a single item. In subsequent iterations, frequent itemsets identified in the previous iteration are extended with one more item to generate larger candidate itemsets. A single scan of the database of expression experiments suffices to determine which candidates generated in an iteration are frequent itemsets. By considering only candidates obtained by enlarging existing frequent itemsets, we greatly reduce the search space of itemsets to be verified. The *a priori* property guarantees that we do not miss any frequent itemsets when using this optimization technique.

Once frequent itemsets are identified, generating association rules from them is straightforward. Any frequent itemset *X* of size greater than one can be divided into two itemsets, *LHS* and *RHS*. The confidence of the rule $LHS \Rightarrow RHS$ is the ratio of the support of *X* and the support of *LHS*. If the confidence of a candidate rule exceeds a specified minimum confidence criterion, the rule is included in the results. In practice, a single frequent itemset can be subdivided into smaller itemsets in a number of ways to generate candidate association rules. In our study using yeast data, we focused on candidate rules where the *LHS* set consisted of a single item (e.g. given the itemset $\{A, B, C, D\}$, check candidate rules such as $\{A\} \Rightarrow \{B, C, D\}$ but not rules such as $\{A, C\} \Rightarrow \{B, D\}$ or $\{A, C, D\} \Rightarrow \{B\}$), this type of rule representing one pattern of interest, though not the only one.

The Apriori algorithm as described above is guaranteed to find all of the frequent itemsets that exist within a data set within a finite amount of time. However, the vast majority of the frequent itemsets found by the algorithm will be redundant in the sense that many of them will actually be subsets of larger frequent itemsets. For example, a single ‘closed’ frequent itemset with 20 items (which may well exist in gene expression data) is made up of $\binom{20}{5} = 15\,504$ frequent itemsets of size 5 and $\binom{20}{10} = 184\,756$ frequent itemsets of size 10. While the analyst may often be more interested in the larger frequent itemsets, building up to the larger itemsets by first working through the intermediate subsets can still take a considerable amount of time. One way to alleviate this problem is to further narrow the search space of candidate itemsets, by specifying additional criteria besides a minimum support in selecting frequent itemsets. For example, in our study of yeast data, where we were interested in generating association rules with the *LHS* set containing a single item, we had our algorithm ignore frequent itemsets that could not form such a rule, since any rule of the above form with *n* items (*n* greater than 2), can be derived by extending a particular rule with *n* – 1 items.

3 IMPLEMENTATION

We developed a database application that implements a version of the Apriori algorithm as described in **Methods** for first finding frequent itemsets and then generating association rules from those itemsets. The application is a Microsoft Access Database Project (ADP), which works by connecting to an SQL Server database. The application is freely available from our web site (see **Abstract**). As input, the application accepts an expression data set in the format of one or more spreadsheets, with items organized by row, and experiments organized by column; each of these spreadsheets is read into a database. The application then mines the database for frequent itemsets that exist

within the data. The application proceeds iteratively using Apriori until all frequent itemsets have been found. The user can also specify some additional criteria besides a minimum support in selecting frequent itemsets of interest, such as requiring selected itemsets to form at least one rule where the *LHS* set has a single item.

Once the data set has been mined for frequent itemsets, the application can then generate association rules from these itemsets. As hundreds of thousands of frequent itemsets may exist in a sizeable data set, the user can have the application limit the search space of candidate rules to those that could be generated from itemsets of a specified size (e.g. all itemsets with more than seven items) or which include at least one item within a specified set of items (e.g. all itemsets that include the genes ‘ADH5’ or ‘LYS1’). To further limit the search space of candidate rules, the application looks only for rules where either the *LHS* or the *RHS* sets of the rule $LHS \Rightarrow RHS$ contain only one item.

Once the frequent itemsets in the database have been mined for association rules, the user can export the results from the database into a spreadsheet. The user can limit the exported results to rules of a specified number of items or which include at least one item within a specified set of items.

4 RESULTS

4.1 Data sets

To demonstrate the algorithm, we used the compendium from Hughes *et al.* (2000) of expression profiles for 6316 transcripts corresponding to 300 diverse mutations and chemical treatments in yeast. We binned an expression value greater than 0.2 for the log base 10 of the fold change as being up; a value less than -0.2, as being down; and a value between -0.2 and 0.2 as being neither up nor down. Of the 6316 transcripts in the data set, 197 were up in at least 10% of the experiments and 47 were down in at least 10% of the experiments; a list of these transcripts can be obtained with our supplementary data for the yeast results (see **Abstract**).

Using the transformed data set, we then constructed a ‘randomized’ data set, which consisted of all of the expression values for each transcript in the original data set being shifted together with respect to the values of the other transcripts by a random number of experiments. The purpose of the randomized data set was to see how many association rules would be found in a data set comparable to the first data set, but one in which the items should not have any relationships between each other.

4.2 Resulting rules

We ran our implementation of the algorithm in two separate cases, using first the randomized data set and then

the yeast data set. In both cases, we specified the minimum support for frequent itemsets to be 10% and the minimum confidence for association rules to be 80%. We specified that all frequent itemsets be able to form at least one rule of the form $LHS \Rightarrow RHS$ (with a confidence of 80%), where the *LHS* set contained a single item. We ran the application on a desktop computer with an Intel Pentium 4 processor. On the yeast data set, the application took about one day to find the frequent itemsets, the longest step in the data mining process (our implementation of Apriori was not a particularly fast one and numerous techniques are described in the data mining literature for making the basic Apriori algorithm run even faster, these techniques often being used on data sets much larger than ours). In the randomized data set, only 1 frequent itemset of size two was found that could form an association rule. Therefore, we can confidently state that practically all of the rules mined from the yeast data set will not have existed by chance.

In the yeast data set, the application found tens of thousands of frequent itemsets of size seven or greater, although this number in itself has little meaning, as the majority of these itemsets are expected to be redundant, i.e. subsets of closed itemsets (see **Methods**). We then did a manual search through the database for those frequent itemsets with seven or more items that appeared to be closed, i.e. itemsets that were not subsets of some larger itemset. From these itemsets, the application generated some 40 rules, with many of the rules being very similar to each other, differing by one or two genes. We list a subset of these rules in Table 1 (the complete list is included with the supplementary data, see **Abstract**). In each of these rules, all of the genes are up; none of the genes happen to be down. To help us put these genes in context with each other, Table 2 gives a description for each of the genes included in a rule in Table 1.

4.3 Interpretation

Rule 1 in Table 1 states that in most (81%) of the cases where the gene YHM1 was up (highly expressed), all of the genes on the right-hand side of the rule were also up. All of the genes involved were up together in 11% of the experiments. The rest of the rules in Table 1 can be interpreted in a similar manner. For rules with a higher number of genes, the support and confidence are typically close to the cutoff threshold, which is why most of the rules in Table 1 have a support and confidence close to 10% and 80%, respectively.

Looking at the rules in Table 1 and the supplementary data, we see a number of genes that are common to many of them: CTF13, HIS5, LYS1, RIB5, SNO1, SNZ1, SRY1, YBR047W, YHR029C, and YOL118C. Individually, these genes have a high support in the data set (around 20–30%), which would help explain their being present in

Table 1. Selected association rules mined from the yeast expression data set of Hughes *et al.* (2000). All of the genes listed in each rule represent the gene being up in the experiment profile. ‘Support’ and ‘Confidence’ give the support and confidence for each rule, respectively

Association rule	Support	Confidence
1 {YHM1} \Rightarrow {ARG1,ARG4,ARO3,CTF13,HIS5,LYS1,RIB5,SNO1,SNZ1,YHR029C,YOL118C}	11%	81%
2 {ARO3} \Rightarrow {ARG1,ARG4,CTF13,HIS5,LYS1,RIB5,SNO1,SNZ1,YHM1,YHR029C,YOL118C}	11%	89%
3 {ORT1} \Rightarrow {ADH5,ARG4,BNA1,CPA2,CTF13,SNO1,SNZ1,YBR047W,YGL117W}	10%	83%
4 {NIT1} \Rightarrow {ATR1,BNA1,CPA2,CTF13,LYS1,RIB5,SNO1,SNZ1,SRY1,YBR047W,YHR029C,YOL118C,YPL033C}	11%	80%
5 {YIL165C} \Rightarrow {ATR1,BNA1,CPA2,CTF13,HIS5,LYS1,NIT1,RIB5,SNO1,SNZ1,SRY1,YBR047W,YHR029C,YOL118C,YPL033C}	10%	81%

many of our rules. The ORFs YBR047W, YHR029C, and YOL118C have not been characterized. The genes HIS5, LYS1, RIB5, and SRY1 are involved in amino acid biosynthesis. SNO1 and SNZ1 are stationary-phase induced genes that appear to be involved in the cellular response to nutrient limitation and growth arrest (Padilla *et al.*, 1998). SNO1 and SNZ1 are both proximal to CTF13 on chromosome 13. CTF13 is a component of the ‘Cbf3’ kinetochore protein complex, which binds to the CDE III element of centromeres during mitosis (Lechner and Ortiz, 1996). The proximity and the co-expression of both SNO1 and SNZ1 with CTF13 lead us to the conjecture that the three genes might be involved in the same biological process.

Looking at rules 1 and 2 in Table 1, we note that YHM1 and ARO3 are found on opposite sides of these rules. YHM1 shares sequence similarity to mitochondrial carrier proteins and has been identified as a multicopy suppressor of an ABF2 mutant lacking the HMG1-like mitochondrial HM protein (Contamine and Picard, 2000; Kao *et al.*, 1996). ABF2, or ARS-binding factor 2, is a mitochondrial protein that plays a possible role in DNA recombination. ABF2 binds specifically to the autonomously replicating sequence ARS1, a likely chromosomal origin of replication (Diffley and Stillman, 1991). ARO3 codes for DAHP synthase, which catalyzes the first step in aromatic amino acid biosynthesis. This step is a major control point of the pathway and synthesis of the enzyme is strongly regulated. The gene ARO3 is activated by ABF1, another ARS-binding factor (Kunzler *et al.*, 1995). Whether the nature of the association suggested here between ARO3 and YHM1 has something to do with the fact that both of these genes have an association with an ARS-binding factor is an open question.

Rule 3 in Table 1 shows a set of genes that are co-expressed with ORT1. The product of ORT1 is a mitochondrial protein that appears to have a role in transporting ornithine from the mitochondria to the cytosol to be further processed into arginine (Crabeel *et al.*, 1996). The genes associated here with ORT1 include ARG4 and CPA2, which code for enzymes that are involved in the synthesis of arginine in the cytosol.

Rule 4 in Table 1 shows a set of genes that are highly expressed with the gene NIT1. Rule 5 shows a set of genes that are highly expressed with the uncharacterized ORF YIL165C. The rules for NIT1 and YIL165C in Table 1 are very similar to each other. YIL165C is homologous and directly adjacent to NIT1 on chromosome 9. The function of the NIT1 gene is not known, but it shares sequence similarity to the NIT1 genes in human, mouse, and *C. elegans*, among others (Pace *et al.*, 2000). The NIT1 genes are members of an uncharacterized gene family with homology to bacterial and plant nitrilases. In human and mouse, NIT1 has been shown to be co-expressed with the FHIT gene. In *C. elegans*, NIT1 and FHIT occur in a fusion protein, NitFhit. (Pekarsky *et al.*, 1998). In humans, FHIT suppresses tumor formation by inducing apoptosis (Pace *et al.*, 2000). Genes that appear in rules with NIT1 include ATR1, which is involved in aminotriazole resistance and is believed to be associated with specific transport systems for effusing hydrophobic drugs (Goffeau *et al.*, 1997), and YPL033C, an uncharacterized ORF that has been shown to be induced in the SOS response in yeast to DNA damaging agents or drugs that inhibit DNA metabolism (Perkins *et al.*, 1999).

5 DISCUSSION

The association rules that we have mined from the yeast data certainly represent only a fraction of all of the possible gene-to-gene interactions that remain to be discovered in yeast. More rules could be found by using different search criteria (e.g. a lower minimum support) or another large data set. The rules that we have found, however, do represent a considerable number of non-random patterns of interest that could lead to the generation of new hypotheses to explain them, hypotheses that could ultimately be confirmed in wet laboratory experiments.

In clustering analysis of expression data, the goal is to define each gene as being part of a self-contained cluster, based on the similarity in the expression pattern of the gene to those of the other genes in the same cluster. Which genes cluster together can vary considerably, both because of the different similarity metrics that can

Table 2. List of the genes included in at least one rule in Table 1

Item	Description	Comment
ADH5	Alcohol dehydrogenase isoenzyme V	
ARG1	Arginosuccinate synthetase	Key enzyme in arginine biosynthesis
ARG4	Argininosuccinate lyase	Key enzyme in arginine biosynthesis
ARO3	DAHP synthase	Catalyzes the first step in aromatic amino acid biosynthesis; strongly regulated; activated by ABF1 (Kunzler <i>et al.</i> , 1995)
ATR1	Aminotriazole resistance	Believed to be associated with specific transport systems for effusing hydrophobic drugs (Goffeau <i>et al.</i> , 1997)
BNA1	Biosynthesis of nicotinic acid	Involved in amino acid synthesis
CPA2	Carbamyl phosphate synthetase	Key enzyme in arginine biosynthesis; subject to general control of amino acid biosynthesis (Messenguy <i>et al.</i> , 1983)
CTF13	Component of the 'Cbf3' kinetochore protein complex, which binds to the CDE III element of centromeres	Involved in mitosis (Lechner and Ortiz, 1996)
HIS5	Histidinol-phosphate aminotransferase	Responsive to control of general amino acid biosynthesis (Nishiwaki <i>et al.</i> , 1987)
LYS1	Saccharopine dehydrogenase	Involved in amino acid biosynthesis
NIT1	Nitrilase	Function unknown; Homologous to NIT1 in mouse, human, <i>C. elegans</i> , which has an association in these organisms with FHIT, a tumor suppressor (Pekarsky <i>et al.</i> , 1998)
ORT1	Mitochondrial integral membrane protein, ornithine transporter	Involved in transporting ornithine from the mitochondria to the cytosol to be further processed into arginine (Crabeel <i>et al.</i> , 1996)
RIB5	Riboflavin biosynthesis	Involved in amino acid biosynthesis
SNO1	SNZ1 proximal ORF, stationary phase induced gene	Proximal to CTF13 on chromosome 13
SNZ1	Snooze: stationary phase-induced gene family	May be involved in cellular response to nutrient limitation and growth arrest (Padilla <i>et al.</i> 1998); proximal to CTF13 on chromosome 13
SRY1	Serine racemase homolog in Yeast	Involved in amino acid biosynthesis
YBR047W	Function unknown	
YGL117W	Function unknown	
YHM1	High copy suppressor of ABF2 ts defect; putative mitochondrial carrier protein	ABF2 is a mitochondrial protein that may play a role in DNA recombination (Contamine and Picard, 2000)
YHR029C	Function unknown	
YIL165C	Function unknown	similar to nitrilases, adjacent to NIT1 on genome; putative pseudogene
YOL118C	Function unknown	
YPL033C	Function unknown	Induced in the SOS response to DNA damaging agents or drugs that inhibit DNA metabolism (Perkins <i>et al.</i> , 1999)

be used to compare any two clusters and because of experimental and biological noise that exists in expression data. Another issue with clustering is that a gene can usually be characterized in more than one way, while it can belong to only one cluster (in hierarchical clustering, we have a hierarchy of clusters within clusters, but a gene cannot belong to two unrelated clusters). Determining the interactions that can exist between different genes is not easily done using clustering results, especially as a gene can participate in more than one gene network. In the case of the Hughes data, the associations that we found between

the genes of interest would not likely have been discovered using hierarchical clustering, as genes that appear to be associated in our rules do not appear adjacent to each other in a clustering of the data (clustering results for the Hughes study (2000) are available with that study's supplementary material).

In contrast, mining expression data for association rules would seem more useful in helping to uncover gene networks. Association rules can describe how the expression of one gene may be associated with the expression of a set of genes. Given that such an association

exists, one might easily infer that the genes involved participate in some type of gene network. However, while an association rule may imply an association, it does not necessarily imply a cause and effect relationship. Determining the precise nature of the association implied by a rule requires prior biological knowledge or further investigation or both. Similar to clustering, one might infer a function for a gene, for which the function is not exactly known, based on the other genes the gene appears with in one or more association rules. Unlike clustering, a gene can belong to any number of rules and is not limited to a single rule. (We, of course, are not suggesting that association rules are 'better' than clustering with respect to gene expression data, only that the two methods are quite different, and that association rules can reveal patterns that might not have been revealed using clustering.)

Association rules can also be used to help relate the expression of genes to their cellular environment. In clustering analysis, one must take care that the values of the clustered elements are all in the same units of measurement. It can be problematic, for example, to combine a binary variable with an expression data set for clustering. In association rules, we can think of items in terms of being up or down, i.e. present or absent. Items in a rule can include the presence or absence of cellular conditions, such as the cells being cancerous or not, the cells receiving a heat shock treatment before profiling or not, etc. For example, association rules could help in the search for 'cancer' genes, especially as the case could exist where no single gene might be responsible for the initiation or progression of cancer, but instead certain sets of genes acting together. Another example of a possible study would be to look for associations between certain attributes of the medical histories of cancer patients and the genes that might be expressed in their corresponding tumors as a result.

Here we applied the standard data mining technique of association rules to gene expression data. In an analysis of a fraction of the rules mined from a data set for yeast, we find numerous associations between certain genes, most of which appear to have biological significance. We plan to use methods of this type in future analyses of biological systems; for example, our group has recently mined an expression data set for breast cancer, finding rules relating clinical outcomes to certain patterns of gene-to-gene associations, the results of which we hope to release soon. In the field of data mining, a vast amount of literature exists on finding association rules, and the techniques that we used here could readily be improved upon and expanded. Our study demonstrates that one can develop database applications that do more than merely store and retrieve expression data, but are tools for doing exploratory data analysis as well.

ACKNOWLEDGEMENTS

The first author was supported in part by a training grant from Pfizer Global Research and Development, Ann Arbor Laboratories. We thank Tom Blackwell, Rork Kuick, George Michailidis, and an unnamed reviewer for their most helpful advice and comments.

REFERENCES

- Contamine,V. and Picard,M. (2000) Maintenance and integrity of the mitochondrial genome: a plethora of nuclear genes in the budding yeast. *Microbiol. Mol. Biol. Rev.*, **64**, 281–315.
- Crabeel,M., Soetens,O., De Rijcke,M., Pratiwi,R. and Pankiewicz,R. (1996) The ARG11 gene of *Saccharomyces cerevisiae* encodes a mitochondrial integral membrane protein required for arginine biosynthesis. *J. Biol. Chem.*, **271**, 25011–25018.
- Doddi,S., Marathe,A., Ravi,S.S. and Torney,D.C. (2001) Discovery of association rules in medical data. *Med. Inform. Internet. Med.*, **26**, 25–33.
- Diffley,J.F. and Stillman,B. (1991) A close relative of the nuclear, chromosomal high-mobility group protein HMG1 in yeast mitochondria. *Proc. Natl Acad. Sci. USA*, **88**, 7864–7868.
- Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
- Friedman,N., Linial,M., Nachman,I. and Pe'er,D. (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.
- Goffeau,A., Park,J., Paulsen,I.T., Jonniaux,J.L., Dinh,T., Mordant,P. and Saier,M.H.Jr. (1997) Multidrug-resistant transport proteins in yeast: complete inventory and phylogenetic characterization of yeast open reading frames with the major facilitator superfamily. *Yeast*, **13**, 43–54.
- Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D., Kidd,M.J., King,A.M. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
- Kao,L.R., Megraw,T.L. and Chae,C.B. (1996) SHM1: a multicopy suppressor of a temperature-sensitive null mutation in the HMG1-like *abf2* gene. *Yeast*, **12**, 1239–1250.
- Kunzler,M., Springer,C. and Braus,G.H. (1995) Activation and repression of the yeast ARO3 gene by global transcription factors. *Mol. Microbiol.*, **15**, 167–178.
- Lechner,J. and Ortiz,J. (1996) The *Saccharomyces cerevisiae* kinetochore. *FEBS Lett.*, **389**, 70–74.
- Messenguy,F., Feller,A., Crabeel,M. and Pierard,A. (1983) Control-mechanisms acting at the transcriptional and post-transcriptional levels are involved in the synthesis of the arginine pathway carbamoylphosphate synthase of yeast. *EMBO J.*, **2**, 1249–1254.
- Nishiwaki,K., Hayashi,N., Irie,S., Chung,D.H., Harashima,S. and Oshima,Y. (1987) Structure of the yeast HIS5 gene responsive to general control of amino acid biosynthesis. *Mol. Gen. Genet.*, **208**, 159–167.
- Pace,H.C., Hodawadkar,S.C., Draganescu,A., Huang,J., Bieganski,P., Pekarsky,Y., Croce,C.M. and Brenner,C.

- (2000) Crystal structure of the worm NitFhit Rosetta Stone protein reveals a Nit tetramer binding two Fhit dimmers. *Curr. Biol.*, **10**, 907–917.
- Padilla,P.A., Fuge,E.K., Crawford,M.E., Errett,A. and Werner-Washburne,M. (1998) The highly conserved, coregulated SNO and SNZ gene families in *Saccharomyces cerevisiae* respond to nutrient limitation. *J. Bacteriol.*, **180**, 5718–5726.
- Pekarsky,Y., Campiglio,M., Siprashvili,Z., Druck,T., Sedkov,Y., Tillib,S., Draganescu,A., Wermuth,P., Rothman,J.H. *et al.* (1998) Nitrilase and Fhit homologs are encoded as fusion proteins in *Drosophila melanogaster* and *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA*, **95**, 8744–8749.
- Perkins,E.L., Sterling,J.F., Hashem,V.I. and Resnick,M.A. (1999) Yeast and human genes that affect the *Escherichia coli* SOS response. *Proc. Natl Acad. Sci. USA*, **96**, 2204–2209.
- Ramakrishnan,R. and Gehrke,J. (2000) *Database Management Systems*. McGraw-Hill, New York, pp. 708–719.
- Stilou,S., Bamidis,P.D., Maglaveras,N. and Pappas,C. (2001) Mining association rules from clinical databases: an intelligent diagnostic process in healthcare. *Medinfo*, **10**, 1399–1403.
- Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E. and Golub,T. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.
- Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet.*, **22**, 281–285.
- Thattai,M. and van Oudenaarden,A. (2001) Intrinsic noise in gene regulatory networks. *Proc. Natl Acad. Sci. USA*, **98**, 8614–8619.