

Mining Association Rules in Microarray Data

John Sikorski

Abstract

Microarray technology arguably has created a revolution in the field of biomedical research, providing researchers with a means of looking at how genes are expressed under certain physical conditions such as drug treatment or disease. Although extremely powerful, significant obstacles stand in the way of fully utilizing this data. Many data mining techniques have been applied to microarray data analysis including hierarchical clustering, k-means clustering, and self organizing maps but very little literature exists on the application of association rules to microarray data. Association rules not only allow us to group similarly expressed genes but also help discern relationships between genes.

Introduction

Up until the recent past, researchers were only able to examine the expression levels of one or a few genes at a time but with the advent of Microarray technology, up to 20,000+ genes can be observed at once. The power of Microarray is in determining relationships between genes, those that are differentially or coordinately expressed under specific conditions. Such genes may provide possible drug target candidates or aide in the understanding of a disease process. Microarray experiments produce vast amounts of data requiring advanced data analysis techniques.

Biology Primer

The central Dogma of Biology states that the flow of information is from genetic material, DNA, to RNA, to protein. The nucleus of each cell contains DNA coding for all proteins found in the body. Depending on its environment, a cell will express only a certain number of genes by translating the double stranded DNA to single stranded RNA through the process of transcription. The RNA is then transferred out of the nucleus of the cell into the cytoplasm where the cells protein production machinery takes over and produces proteins from the RNA blueprints through the process of translation. A cell can regulate the amount of protein present by up-regulating or down-regulating the production of RNA which inevitably increases or decrease the amount of protein produced.

Proteins are responsible for almost all biological processes in the body. Disease states can often be traced back to a change, for some reason, in the levels of certain proteins. Biotechnology drug development frequently tries to understand and target these protein interaction pathways. Microarray technology is able to look at the expression levels of thousands of genes at once.

Microarray Process Primer

A DNA microarray consists of a glass or nylon slide coated with many spots, each of which contains many identical DNA sequences known as probes. Each probe is a piece of single stranded complimentary DNA (cDNA) to a gene of interest. A researcher

typically looks at expression levels in certain tissues. The tissue of interest is dissected from a drug treated, diseased, or control animal and RNA is extracted from it. Complimentary copies of all RNA present are produced and labeled with fluorescent dyes, producing labeled cDNA. A solution containing the labeled cDNA is introduced to the microarray chip. Through the base pairing principle described by Watson and Crick, the labeled cDNA binds to the cDNA on the chip with A binding to T and C binding to G. Therefore, the sequence AGTCTA would only bind to TCAGAT. Unbound cDNA is washed from the slide and the relative amounts of bound cDNA are determined by measuring the fluorescence of each spot. A series of data cleaning and normalization steps produces numerical data representing the expression levels of each gene compared to control. A value of 1 means no change, less than one means the gene is under expressed, and greater than one means the gene is over expressed (see Table 1).

Array	Gene				
	A	B	C	D	E
1	0.9	2.6	0.4	5.5	1.1
2	3	0.1	2.4	0.7	1.2
3	5.6	0.3	9.5	0.09	3.8
4	0.3	8.9	2.6	1.4	1.4
5	9.2	5.4	1.6	0.8	0.2
6	2.1	0.7	15.2	1.2	6.3
7	0.02	4.8	3.2	0.3	0.5
8	0.6	0.6	0.4	0.5	1.6
9	3.3	2.6	0.3	1.1	0.8
10	0.9	2.1	4.1	0.9	1

Table 1: Microarray data showing relative gene expression levels. Value of 1 is no change, less than one is under expressed, and greater than one is over expressed.

Common Microarray Data Analysis Techniques

Hierarchical clustering (Eisen *et al.* 1998) is probably the most extensively used microarray data mining technique, using one of several techniques to iteratively, starting with one gene, combine genes with their nearest neighbor, gradually building clusters and associations of clusters, resulting in a hierarchical tree (Figure 1). Distance between clusters is defined by the distance between their average expression patterns. A visual representation of the clusters is created in the form of a hierarchical tree, or dendrogram, familiar and easily understood by all biologists. The tree structure makes it easy to visually see how similar the expression patterns are between genes or sets of genes. Hierarchical clustering makes a very good ‘first pass’ over the data.

Non-hierarchical clustering techniques group N number of genes into K clusters. Two examples are K-Means clustering (Tavazoie *et al.*, 1999) and Self Organizing Maps (Tamayo, *et al.*, 1998). K-means uses a predefined (K) number of clusters, or ‘centroids’. Using a three step process, genes are randomly assigned to a centroid. The mean inter- and intra-cluster distances are then calculated. The third step moves genes from one cluster to another. Steps two and three are repeated until intra-cluster distance is minimized and inter-cluster distance is maximized, typically resulting in K, round shaped clusters. K-means excels at clustering similarly expressed genes. For example, a dataset

containing cancerous and non-cancerous tissues could use K-means to identify 2 groups of genes, those that change with cancer and those that don't.

Self Organizing Maps (SOM) (Figure 2) use neural network techniques to iteratively map nodes into n-dimensional 'gene expression space'. First, random vectors are created and added to each node. Next, the distance between the vectors and a randomly selected gene are calculated. The vector closest to the gene is updated, making it more like the genes vector. The process is repeated thousands of times until no more changes can be made. SOM have characteristics making them well suited for microarray data analysis. By their nature, large dimensional gene space is converted into something more manageable and understandable. SOM's also allow for the use of prior knowledge by imposing a partial structure (number of clusters and dimensionality) on the analysis.

All clustering techniques excel under certain conditions but all have their drawbacks. Hierarchical clustering imposes a rigid relational structure on the data which may or may not reflect reality. K-means and SOM requires a predetermined number of clusters. This works well in certain situations, like the cancer example above, but for blind, exploratory data analysis, like determining gene relationships, the number of clusters will not be known ahead of time. K-means has an additional drawback in that it produces fairly round clusters, resulting in inaccurate identification of close or geometrically shaped clusters.

Genes frequently interact with several different pathways but all clustering techniques suffer from the fact that a gene can only be a member of one and only one cluster. K-means does show relationships up the tree but a gene cannot be a member of a cluster on an opposite branch. Lastly, although clustering shows an association between groups of genes, no conclusions can be drawn about relationships between genes within a cluster, such as a direction of action. Association rules help us to find relationship between genes, relationships between a gene and several other groups of genes, and possibly provide a direction of action.

Association Rules

Widely used in the area of 'market basket analysis', association rules take the form of LHS \rightarrow RHS where RHS and LHS are both sets with RHS likely to occur whenever LHS occurs. For example, a stores sales database can be mined looking for relationships for what people buy when they also purchase milk. Association rules such as {Milk} \rightarrow {Cereal, Bisquick} and {Milk} \rightarrow {Bisquick, Maple Syrup} may be uncovered. Such obvious associations are not necessarily very useful or insightful but by altering user defined settings, not so obvious associations may be realized.

Association rules can also be applied to microarray data (Creighton & Hanash, 2003) and (Becquet, *et al.*, 2002). Instead of looking at contents of shopping carts, as with market basket analysis, we can look at relationships between genes in microarray experiments; treat the genes as the "items" and the arrays as the "transactions".

Association rules applied to Microarray Data

Gene expression data requires a few steps of data processing before it can be analyzed for association rules. In market basket analysis an item is either purchased or not purchased but microarray data consists of continuous numerical data. The first step is to discretize the data, convert it to a Boolean or tertiary notation. We define a cutoff

value where anything above this setting will be considered up regulated, assigned a value of '1', and anything below will be considered down regulated, assigned a value of '0'. If we want to look at three conditions, up regulated, down regulated and unchanged we could use the values of 1,-1, and 0. The data in Table 1 was discretized into Boolean notation to produce Table 2. Although any value greater than 1 can be considered up regulated, a cutoff of 2 was chosen to avoid inherent experimental noise and error.

Array	Gene				
	A	B	C	D	E
1	0	1	0	1	0
2	1	0	1	0	0
3	1	0	1	0	1
4	0	1	1	0	0
5	1	1	0	0	0
6	1	0	1	0	1
7	0	1	1	0	0
8	0	0	0	0	0
9	1	1	0	0	0
10	0	1	1	0	0

Table 2: Discretized data from Table 1 in Boolean notation. Values ≥ 2 were assigned '1' and values < 2 were assigned 0.

Association rules require two user defined values, coverage and confidence. Coverage defines how often an item occurs in the dataset, in our case, how often a gene is up regulated compared to the total number of experiments:

$$\text{Coverage}(\text{Gene}(x)) = (\text{count Gene}(x) \uparrow) / (\text{count arrays})$$

Using the notation $\text{LHS} \rightarrow \text{RHS}$, confidence describes the likelihood that RHS is present whenever LHS is present. Therefore,

$$\text{Confidence} = (\text{count LHS} \uparrow \text{ and RHS} \uparrow) / (\text{count LHS} \uparrow)$$

Or written another way:

$$\text{Confidence} = (\text{coverage}(\text{LHS} \uparrow \text{ and RHS} \uparrow) / \text{coverage}(\text{LHS} \uparrow))$$

The first step is to find itemsets where all items in the set meet the coverage cutoff. It becomes obvious that for even a small dataset the number of possible itemsets is very large, growing exponentially with the number of columns. Microarray datasets typically have thousands of columns so techniques must be used to reduce the number of comparisons, such as the Apriori algorithm. The Apriori algorithm relies on a simple property that all subsets of frequent itemsets must also be frequent itemsets. The algorithm proceed iteratively, first finding frequent itemsets of single genes, then adding genes to the sets and removing sets that do not meet the coverage cutoff.

With a coverage set to 0.3, we can see genes A, B, and C meet our cutoff, having coverage's of 0.5, 0.6, and 0.6 respectively. Genes D and E fail with coverage's of 0.1 and 0.2. Knowing this, we just greatly reduced the number of sets to consider since we can ignore genes D and E. We now have three sets containing a single gene each. Each set is expanded by adding another single gene to create {A,B}, {A,C}, and {B,C}, avoiding duplicate genes in each set and duplicate sets. Next we check the coverage of

both genes together in each itemset to find 0.2, 0.3 and 0.3 respectively. Itemsets {A,C} and {B, C} pass our test. If we had more columns, we could continue to iteratively add genes and check the coverage. To see if we have a rule with {A, C} we first check the confidence of $\{A\} \rightarrow \{C\}$. We find previously that A is up in 5 arrays, of these arrays, we find C is up in 3, giving us a confidence of $3/5=0.6$. If we have our confidence set to 0.6 or less, this would pass as a rule. Next we can check the rule in the opposite order, $\{B\} \rightarrow \{A\}$. Here we find a confidence of $3/6=0.5$ so with a confidence set to 0.6 this would fail. The itemset {A, C} produces one rule, $\{A\} \rightarrow \{C\}$, with our coverage and confidence settings.

Results

The microarray data consists of 23 array experiments with 8,011 genes each. Two control sets were produced by shuffling the data. The first set contained completely shuffled data with cells from different rows and columns interchanged. The second set was shuffled by only interchanging cells from within each row.

Custom MatLab code was written to mine the association rules by first determining frequent itemsets with coverage of 0.3. Since a very large number of itemsets and association rules are possible, I stopped generating itemsets once 7 items per set was reached (Table 3). Association rules in the form of LHS \rightarrow RHS were determined using a confidence cutoff of 0.75 (Table 4). For simplicity, I considered only rules where LHS consisted of one gene and RHS the remaining genes in the set. This also produces rules that answer a more frequently asked question; what genes are co expressed with my gene of interest? Mining the control data sets yielded no valid itemsets even at an itemset size of one, no genes met our coverage cutoff.

Number of Itemsets			
Itemset Size	Microarray	Control 1	Control 2
1	31	0	0
2	149	0	0
3	387	0	0
4	564	0	0
5	666	0	0
6	301	0	0
7	137	0	0

Table 3: Number of valid itemsets for each itemset size. Control 1 is completely shuffled data and control 2 was shuffled by row only.

Number of Association Rules			
Itemset Size	Microarray	Control 1	Control 2
1	31	0	0
2	195	0	0
3	467	0	0
4	666	0	0
5	680	0	0
6	556	0	0
7	345	0	0

Table 4: Number of valid association rules for each itemset size.

Discussion

Thousands of potential association rules were uncovered from this dataset while the control sets both produced no valid itemsets or rules. The control sets were meant to match the expression values of the real dataset while shuffling the gene/expression associations. Two separate sets were used to control for the possibility of the completely shuffled set being skewed since control arrays were mixed with experimental arrays. Regardless, no association rules were determined from the shuffled data. This suggests there is order to the microarray data and this order was detected by the association rules implementation. The mined rules did not occur by chance alone.

A number of these rules are redundant, being subsets of larger sets. The counts were monitored at each stage to view the progress of the algorithm not to collect all possible rules. To completely mine the data, more itemsets should be produced and many more rules considered. It is interesting to see the itemset size and valid rules peak at an itemset size of 5. Could this have biological significance? Genes typically interact with a small number of other genes, typically in the neighborhood of 5-10 so this observation may hint at real gene associations are being detected. The dataset used for this paper is admittedly small. In order to properly mine for association rules a significantly larger dataset should be used, on the order of hundreds rather than tens of arrays.

Although we were able to find order in the data, the biological relevance of the rules still needs to be determined. A brief review of the top rules showed biological significance in the gene associations (an observation, data not shown). A closer study of many rules needs to be performed to properly assess the value of this technique.

If we assume for the moment that these rules have biological significance, we have found associations that may not have been possible with other common microarray data mining techniques. As mentioned earlier, many data mining techniques have been used, the most common being forms of clustering. The techniques applied by Eisen *et al.* (1998), Tavazoie *et al.*, (1999), and Tamayo *et al.* (1998), hierarchical clustering and self organizing maps, both group genes with similar expression patterns but neither are able to show relationships between sets of genes. In addition, all clustering techniques force a gene to be a member of a single cluster, negating the associations of genes with influences on disparate pathways. Association rules help us pull out these subtle relationships.

Very little literature exists on the application of association rules to microarray data. Creighton & Hanash (2003) and Becquet *et al.*, (2002) both published similar papers on the application of association rules to gene expression data. Both claim similar success as to what has been described in this paper. Association rules are a potentially useful tool in mining gene expression data. I can only surmise the lack of literature on the subject is because microarray data analysis is still in its infancy. The technology has been in use for only about 6-8 years with large numbers of experiments just becoming available for analysis. Association rules will likely play a larger role in microarray data analysis in the future, especially in deciphering gene networks.

References

- Becquet, C., Blachon, S., Jeudy, B., Boulicaut, J., Grandrillon, O. (2002) Strong association rule mining for large scale gene-expression data analysis: a case study on human SAGE data. *Genome Biology*, **12**, 1-16.
- Creighton, C., Samir, H. (2003) Mining gene expression databases for association rules. *Bioinformatics*, **19**, 79-86.
- D'haeseleer, P., Liang, S., and Somogyi, R. (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, **16**, 707-726.
- 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.
- Ramoni, M., Sebastiani, P., and Kohane, I. (2002) Cluster analysis of gene expression dynamics. *PNAS*, **99**, 9121-9126.
- Tamayo, P., Slonim, D., Mesirov, J., Zho, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. and Golub, T. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and applications to hematopoietic differentiation. *Proc. Natl. Acad Sci USA*, **96**, 2907-2912.
- Tavazoi, S., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, G.M. (1999) Systematic determination of genetic network architecture. *Nature Genetics*, **22**, 281-285.

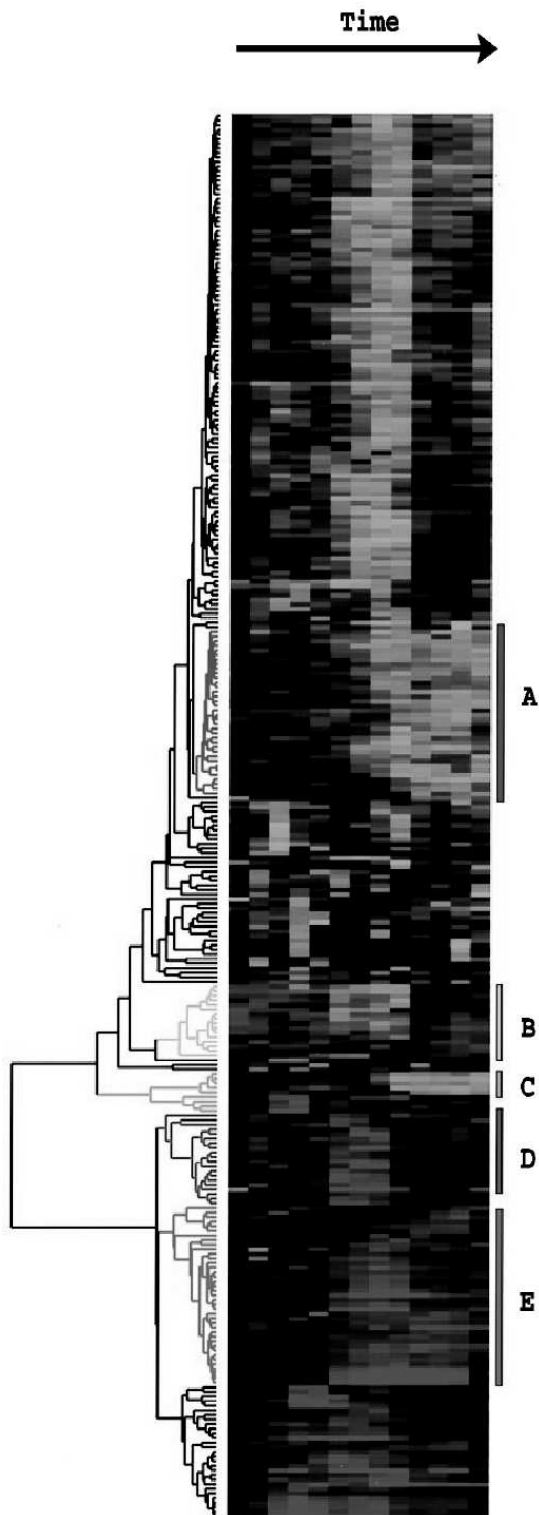


Figure 1: Example of hierarchical clustering (from Eisen et al (1998)). Each row is a gene and each column is an array. Colors represent expression intensities, red overexpressed and green underexpressed. Black signifies no change.

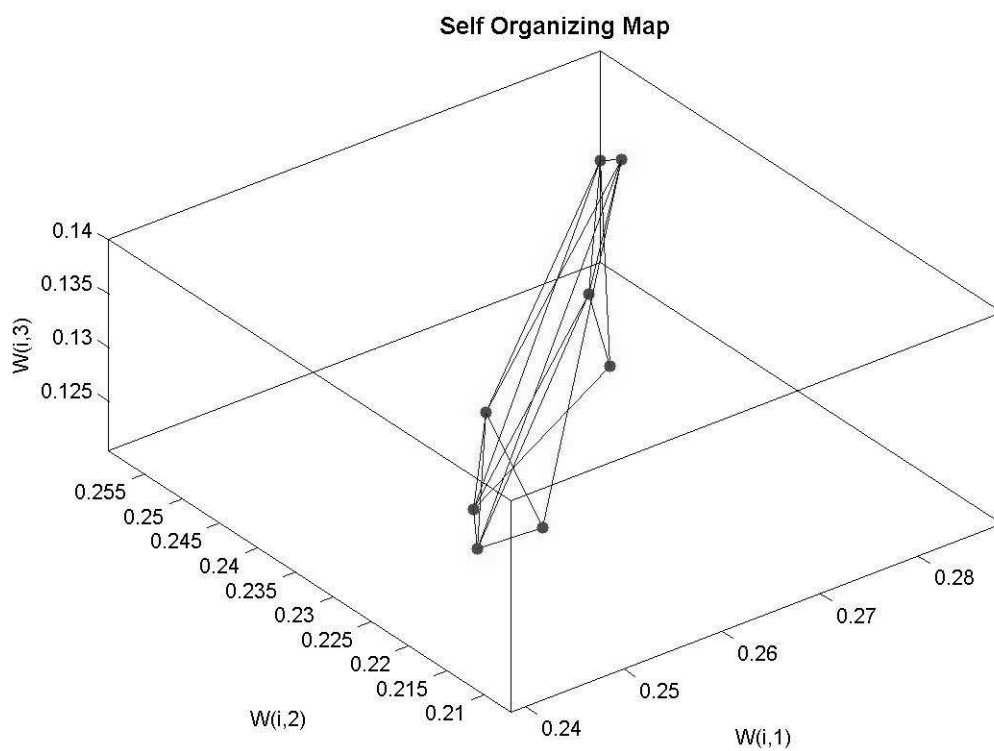


Figure 2: Graph of nodes for self organizing map run on the microarray data in this paper using a [2,2,2] structure..