ST06

Gene Expression Profiling of DNA Microarray Data using Association rule and structural equation modeling

Mussie Tesfamicael

Abstract

The purpose of this study is to use structural equation modeling and association rules to extract meaningful information from bioinformatics data for the purpose of constructing gene networks. Structural equation modeling is widely used in the social sciences to model cause and effect relationships, while association rules are used widely in the area of retail marketing to find items that are related. Both of these methods can be applied to analyze microarray expression data as well.

Structural equation modeling in SAS/Stat, or SEM (PROC CALIS), explains a cause and effect relationship between variables, in this case genes. Association Rules in Enterprise Miner are useful to determine an important set of rules for a dataset with high values of support and confidence.

Association rules will determine the direction of the association between gene markers, so that differentially expressed genes can be analyzed using SEM. The associations between genes can help illustrate how a particular gene is affected by other genes. The associations between genes and categories can describe what genes are expressed as a result of certain cellular environments.

INTRODUCTION

High-throughput technology has changed the dimension of biotechnology. In order to understand the expression level of a gene, one has to determine how the expression level of a certain gene might affect the expression level of other genes; the genes of interest could be on the same cluster or on the same network. By a gene cluster, we mean a set of genes grouped on the same class, while by gene network; we mean a set of genes being expressed together in a non-random pattern [1]. Many data mining techniques have been applied to microarray data analysis, including k-means clustering, hierarchical clustering, self-organizing maps (SOMs), and support vector machines, but very few papers exist in the subject literature that use the concept of association rules and structural equation modeling to extract differentially expressed genes. Recently, [4] used the notion of structural equation modeling to find the causal model of the genes.

Structural equation modeling, or SEM, is a very general, chiefly linear, chiefly cross-sectional statistical modeling technique. It explains a cause and effect relationship between variables, in this case genes, but the limitation of this method is that it doesn't determine the direction of the cause and effect relationship. Association Rule (market basket analysis) is a data mining technique that is useful to determine an important rule for a data set with high values of support and confidence. 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. In market basket analysis, an association rule represents a set of items that are likely to be purchased together; for example, the rule {bag of tortilla} $\Rightarrow$ {jar of salsa} would state that whenever a customer purchases a bag of tortilla, he is likely to purchase a jar of salsa. When the Association rule is applied to a gene expression data set, the item sets represent the genes that are differentially expressed as a result of the other genes being differentially expressed.

BACKGROUND

The associations between genes can help illustrate how a particular gene is affected by other genes. The associations between genes and categories can describe what genes are expressed as a result of certain cellular environments (e.g., a cancer cell). The dataset, yeast, was obtained from Amada software [3].A yeast dataset with 200 genes at 17 different time points was analyzed. Here, the main aim is to find which of the genes are in the same network and are differentially (similarly) expressed as a result of the other genes.

Association rules

An association rule has two numbers that express the degree of uncertainty, beside the antecedent (LHS) and consequent (RHS). The LHS and RHS are sets of items called item sets that are disjoint (LHS $\cap$ RHS = ø). The LHS is the support for the rule. It is the number of times that the combination appears. The support is the number of transactions that include all items in the LHS and number of transactions that include all items in RHS as well as the LHS to the number of transactions that include all items in the RHS. The association node in SAS/STAT requires one identification input (time point), target (gene) and id (gene sequence) [2]. An association rule has the form LHS $\Rightarrow$ RHS, where LHS and RHS are itemsets. Itemsets can be genes in microarray data or transactions in industry. The following will explain the terms used in association rule:

Expected confidence is the percentage of times the item RHS occurs in the data.

Confidence is the percentage of cases in which the item RHS is present when item LHS is present.

Support is the percentage of records containing both item RHS and item LHS.

Lift is how much more likely item RHS is if item LHS happens. A rule has lift when its confidence is higher than its expected confidence.

Count is the frequency of items LHS and RHS occurring together.

STRUCTURAL EQUATION MODELING

The structural equation model is of the form:

$$\begin{matrix} Effect \\ Variable \end{matrix} = \text{Sum} \begin{matrix} Structural \\ Coefficient \end{matrix} \times \begin{matrix} Casual \\ Variable \end{matrix} + \text{Disturbance}$$

such that

$$\underset{(p\times 1)}{Y_i} = \underset{(p\times m)(m\times 1)}{\Lambda_{i1} X_1} + \underset{(p\times m)(m\times 1)}{\Lambda_{i2} X_2} + \ldots + \underset{(p\times m)(m\times 1)}{\Lambda_{ip} X_p} + \underset{(p\times 1)}{\varepsilon_i}$$

where the $\mathbf{Y_i}$ and the $\mathbf{X_I}$**'s** are the names of the gene expressions. The $\mathbf{\Lambda_{i1}}$**'s** are the estimated paths for the genes and the $\boldsymbol{\varepsilon_I}$ are the error terms (Disturbance). The model contains two types of variables, namely exogenous and endogenous variables. Exogenous constructs are independent variables in all equations in which they appear; while endogenous constructs are dependent variables in at least one equation-although they may be independent variables in other equations in the system.

PROGRAM STATEMENT & DATA PROCESSING

The yeast data has 200 genes as rows and 17 time points as columns. The PROC CALIS program analyzes a simple recursive path model. Before writing the program, it is better to visualize the pictorial diagram of how the variables (genes) are related to each other. The association rule is used to find which of the genes are associated with one another. The correlation between all the genes is computed and a bench mark of 0.95 is selected to find the pathway analysis between genes. The correlation matrix is input in the data statement CORR, and the LINQES is based on this correlation matrix to find which of the variables can be written as a linear function of a given antecedent variable. In the microarray notation, it means which of the genes are up/down expressed as a result of a given gene being up/down expressed. A gene is measured "up " (highly expressed) and gene is measured "down" (highly repressed).

```
DATA MICROARRAY
Input
       Gene1,gene2,…,genek; /* Here the variables, V1, V2,…,V13 represent
the genes
CARDS;
```

```
CORR
```

The data are input into a correlation matrix obtained among the thirteen genes.

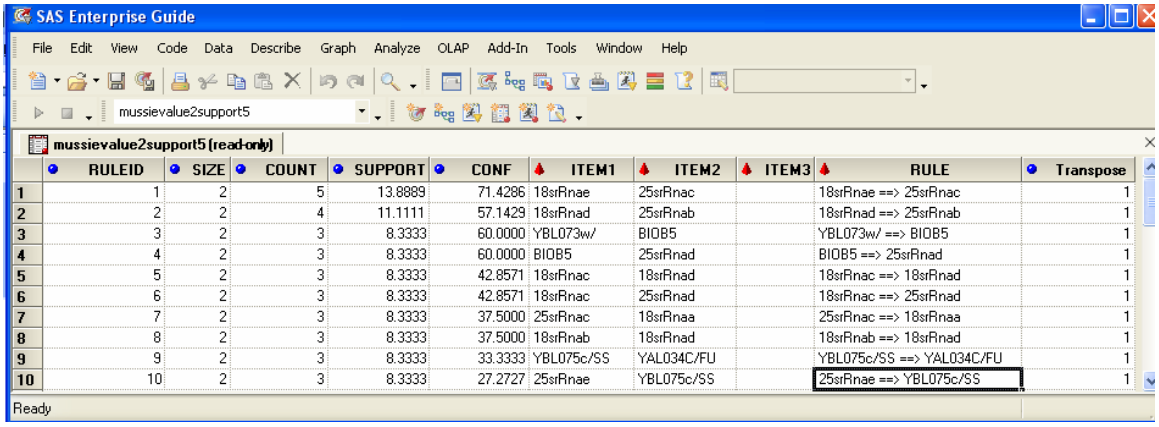
```
PROC CALIS COVARIANCE CORR RESIDUAL MODIFICATION;
LINEQS
        V1=PV1V3 V3  + E1,
        V2=PV2V4 V4  + PV2V5 V5  + E2,
STD
        E1   = VARE1,
        E2   = VARE2,
        V3 =   VARV3,
        V4 =   VARV4,
        V5  = VARV5;
COV
        V6 V9    = CV6V9,
        V9 V12   = CV9V12,
VAR V1 V2 V3 V4 V5 V6 V9 V12;
RUN;
```

The **LINEQS** statement makes the equations that determine the analysis of the gene markers for the data being studied; each line consists of the variables, the path coefficients and the error term. The **STD** statement determines the variance of the endogenous variables (Independent) and the error term of the exogenous variables (independent). The COV statement determines the covariance among each pair of endogenous variables (dependent).

Methodology

The association rules describe how the expression of one gene is associated with the expression of a set of genes. The association generated by association rule suggests genes that are involved in forming a gene network. However, even though association rules imply an association, they do not necessarily imply a cause and effect relationship. The yeast data consists of 200 genes as rows and 17 time points as columns. In order to get important associations between the genes, the data were filtered into three columns with the gene as one column, a sequence id, and the expression of each gene (time point). SAS/STAT Enterprise Miner, version 5, was used to find associations between the genes.

A support of 2.7 or more generated 291 rules, in which case, some of the rules might be redundant, while a support of 8 ore higher gave only 10 rules. The genes with the higher support (>8) and confidence were chosen to be analyzed. In fact, these genes were statistically significant when a t-test was performed to see if a gene's expression level differs from a sequence of time points (0,10,…,160). Once the direction of association between the genes was determined, structural equation modeling (SEM) was applied. The drawback with SEM is that the direction of cause is not determined, but once we get the direction of association between the genes using SAS/STAT and Enterprise Miner, we can apply the method of structural equation modeling (SEM) to test the hypothesis that the path coefficient is zero, meaning that there is no relationship between the exogenous variables and endogenous variables.

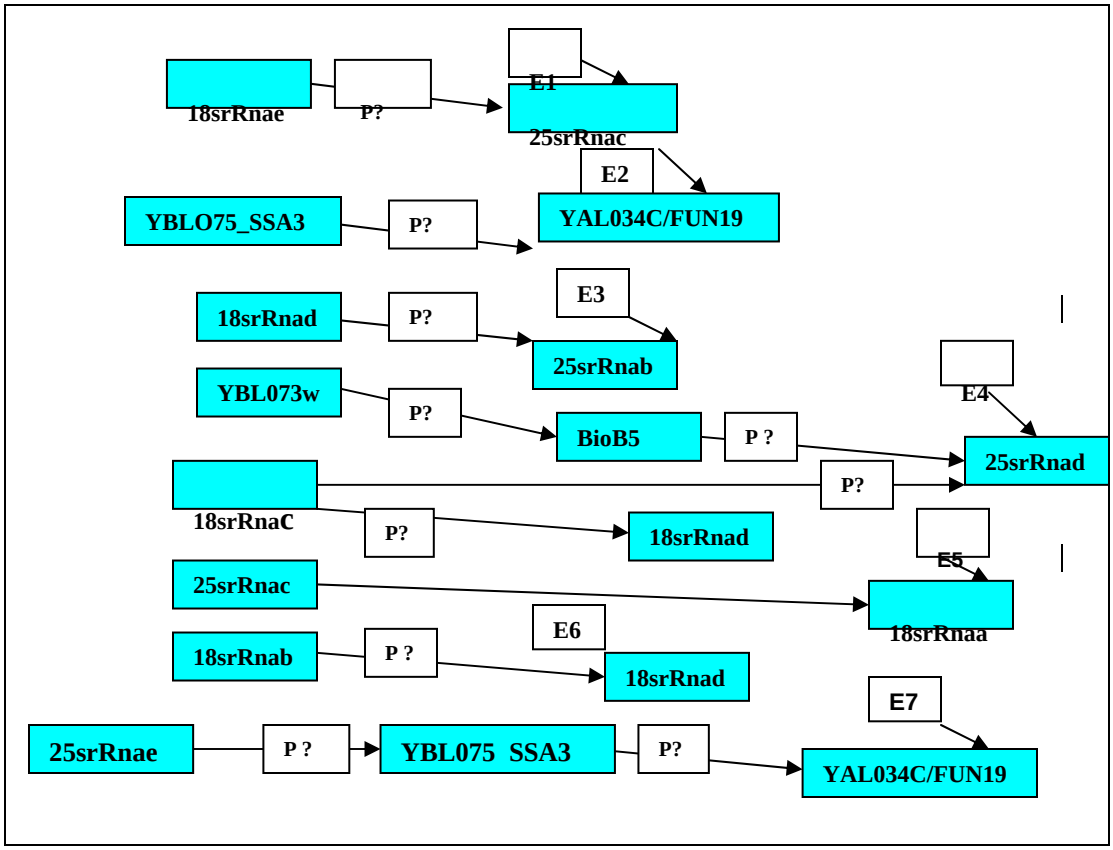


	RULEID	SIZE	COUNT	SUPPORT	CONF	ITEM1	ITEM2	ITEM3	RULE	Transpose
1	1	2	5	13.8889	71.4286	18srRnae	25srRnac		18srRnae ==> 25srRnac	1
2	2	2	4	11.1111	57.1429	18srRnad	25srRnab		18srRnad ==> 25srRnab	1
3	3	2	3	8.3333	60.0000	YBL073w/	BIOB5		YBL073w/ ==> BIOB5	1
4	4	2	3	8.3333	60.0000	BIOB5	25srRnad		BIOB5 ==> 25srRnad	1
5	5	2	3	8.3333	42.8571	18srRnac	18srRnad		18srRnac ==> 18srRnad	1
6	6	2	3	8.3333	42.8571	18srRnac	25srRnad		18srRnac ==> 25srRnad	1
7	7	2	3	8.3333	37.5000	25srRnac	18srRnaa		25srRnac ==> 18srRnaa	1
8	8	2	3	8.3333	37.5000	18srRnab	18srRnad		18srRnab ==> 18srRnad	1
9	9	2	3	8.3333	33.3333	YBL075c/SS	YAL034C/FU		YBL075c/SS ==> YAL034C/FU	1
10	10	2	3	8.3333	27.2727	25srRnae	YBL075c/SS		25srRnae ==> YBL075c/SS	1

Figure 1. The association rule generated with support > 8

The item1 is the antecedent (LHS) and Item2 and item3 are consequents (RHS) of the rule.
Rule 1 in Figure1 states that in most (71.42889 %) of the cases where the gene18srRnae was differentially expressed (up/down), the gene on the right hand side, 25srRnac, is also differentially expressed (up/down). The rest of the rules in Figure 1 can be interpreted in the same manner

Figure 2. Graphical representation of a structural equation model for Gene Markers



The **P** values listed in Figure 2 represent coefficients that relate the gene markers, and **E** represents residuals. Other equations are defined similarly. The correlations are computed and entered in a data statement. The variable on which an arrow is pointed towards is an endogenous variable, which occurs as a result of the gene on which the arrow is coming from. Gene 18srRnad being differentially expressed causes gene 25srRnab to be differentially expressed.

Results

The initial model gave a large chi-square of 211.8154 with df = 54, p= <.0001. Although the chi-square test is a useful index, it is generally accepted that it should be interpreted with caution and supplemented with other goodness of fit indices. This is because the Chi-square test can be influenced by factors in addition to the validity of the theoretical model; these factors include departures from multivariate normality, sample size, and even the complexity of the model. The SAS/STAT Users Guide says the chi-square test statistics provides a “test of the specified model vs. the alternative that the data are from a multivariate normal distribution with unconstrained covariance normal distribution with unconstrained covariance matrix” (SAS/STAT Users guide 1989, volume 1, p. 139). Bentler and Bonett’s (1980) normed-fit index (NFI) has been proposed as an alternative to the chi-square test. Values on this index may range from 0 to 1, with values over 0.9 indicative of an acceptable fit of the model to the data. This index may be viewed “as the percentage of observed-measure covariation explained by a given measurement or structural model (compared with an overall, null model that solely accounts for the observed measure variances)” (Anderson and Gerbing, 1988, p. 421). Although the NFI has the advantage of being easily interpreted, it has the disadvantage of sometimes underestimating goodness of fit in small samples.

A variation on the NFI is the non-normed fit index (NNFI, Bentler & Bonett, 1980). The NNFI has been shown to better reflect model fit at all sample sizes (Bentler, 1989; Anderson & Gerbing, 1988; Marsha, Balla, & McDonald, 1988). NNFI values over 0.9 are also viewed as desirable, although, unlike the NFI, the NNFI may assume values below 0 above 1.

Bentler's (1989) comparative fit index (CFI) is similar to the NNFI in that it provides an accurate assessment of fit regardless of sample size. In addition, the CFI tends to be more precise than the NNFI in describing comparative model fit (Bentler, 1989). Values of the CFI will always lie between 0 and 1, with values over 0.9 indicating a relatively good fit.

The correlation matrix was used on the data statement; the standardized path coefficients are tested to determine whether the path is significant or not. The null hypothesis to be tested in SEM is whether the path coefficient is zero, meaning that there is no relationship between the exogenous (RHS) variables and endogenous variables (LHS). One of the characteristics of an ideal fit is that the absolute values of entries in the normalized residual matrix should not exceed 2 (for the model1 fitted, some of the genes have marginal chi-square values); the p-values associated with the model Chi-square test should exceed 0.05, for this initial model, it is<0.001; the comparative fit index (CFI) and the non-normed fit index (NNFI) should be relatively large (>0.9), but for the gene model1 fitted, it was 0.6662. Hence, the model1 needs to be modified, i.e. some of the paths have to be added or removed based on these diagnostic values, and the standardized path coefficients (less than 0.05 means remove the path and greater than 0.05 means keep the path). Technically removing a path or adding a path is creating the gene network, finding the genes associated with a particular gene.

To find the perfect fit of the model of gene network for the yeast data set, several models are constructed at each step, improving the performance from the previous fit. The analysis is based on getting a non-significant chi-square, a non-normed fit index (NNFI), a normed index (NI), and a comparative fit index (CFI) to have a value of greater than 0.9; also, one has to check to make sure that the correlation between the exogenous variables (genes in this case) is reasonably high. The researcher has to check for significant Lagrange multiplier gamma indices and Lagrange multiplier beta indices; in this case for those gene interactions with significant values, indicating the addition of that specific path.

We have gone through these several stages to find the gene network and the association between several genes. The notion of association rule in this study gives the starting model, the direction of the association between the genes, as structural equation modeling (SEM) lacks this property.

Table 1.
Goodness of Fit Indices for various Models, Genes model study

Model	Chi-square	df	p	NFI	NNFI	CFI
Model1	211.8154	58	<0.0001	0.6662	0.6283	0.7236
Model2	195.0221	57	<0.0001	0.6926	0.6606	0.7520
Model3	152.9313	48	<0.0001	0.7590	0.6936	0.8115
Model4	96.4150	46	<0.0001	0.8481	0.8464	0.9094
Model5	94.1530	45	<0.0001	0.8516	0.8469	0.9117
Model6	88.7769	46	0.0002	0.8601	0.8697	0.9231
Model7	84.2085	45	0.0004	0.8673	0.8779	0.9295
Model8	81.7830	44	0.0005	0.8711	0.8796	0.9321
Model9	81.6809	43	0.0003	0.8713	0.8739	0.9305
Model10	75.5819	42	0.0011	0.8809	0.8879	0.9397
Model11	66.1451	41	0.0077	0.8958	0.9140	0.9548
Model12	53.6509	40	0.0736	0.9123	0.9452	0.9719

Model12 in table1 is chosen as the final model for the association of the gene network of the yeast data, as this model has a non significant p-value 0.0736, Normed fit index (NFI), Non-normed fit index and comparative fit index value greater than 0.9.

The CALIS Procedure
Covariance Structure Analysis: Maximum Likelihood Estimation

Manifest Variable Equations with Standardized Estimates

V1 = -0.0591*V2 + 0.6467*V5 + -0.1735*V6 + -1.3602*V10 + -8.8528*V11
PV1V2 PV1V5 PV1V6 PV1V10 PV1V11

+ 2.4005*V13 + 6.6060*V9 + 0.5111 E1
PV1V13 PV1V9

V2 = -0.7772*V1 + -0.4784*V8 + 1.7176*V9 + -0.9675*V12 + 0.3954 E2
PV2V1 PV2V8 PV2V9 PV2V12

V5 = 0.2305*V8 + 0.8144*V9 + 0.2471 E3
PV5V8 PV5V9

V6 = 0.4859*V1 + 0.8875 E4
PV6V1

V8 = 0.0296*V1 + 3.4749*V2 + -0.2315*V13 + 2.5478*V3 + 0.6224 E5
PV8V1 PV8V2 PV8V13 PV8V3

V10 = 0.9498*V8 + 3.7828*V11 + -3.5320*V4 + 0.3656 E6
PV10V8 PV10V11 PV10V4

V11 = 0.9304*V5 + 1.5480*V10 + 0.9083*V4 + -2.3231*V7 + 0.2979 E7
PV11V5 PV11V10 PV11V4 PV11V7

V13 = -0.0570*V2 + 0.8755*V5 + 0.2132*V8 + 0.2142 E8
PV13V2 PV13V5 PV13V8

Figure 4. Path coefficients of the genes

The independent variables (V2,V5,V6,V10 and V11) with path coefficients (-0.0591,0.6467,-0.1735,-1.3602,-8.8528) are used respectively in the prediction of V1. In the same way, the other path coefficients can be explained.

Discussion

A gene might be associated with several genes, as was found by the association rule; but hierarchically, a gene can not be clustered to several clusters. The main aim of combining SEM with association rules is to see what genes are associated with what genes; so that when a particular gene is expressed up/down, we see what effect it has on the expression level of the other gene. We can infer that the higher the support and confidence for a certain rule, the higher the probability that if a gene has its expression up/down (LHS), all the genes on the right hand side of the rule will have expressed (up/down) as well.

Conclusion

Association rules show which of the genes are associated with each other when they satisfy a certain pre-specified support and confidence. Once we know the direction of association between the genes, we can apply structural equation modeling to predict the paths for the genes. Several models have to be analyzed to get the final model to get a non-significant p-value and higher NNFI,NFI, CFI values that exceed 0.9 as well as a normalized residual for the genes with a value less than 2. The notion of combining association rules and structural equation modeling will give biologists the ability to study the direction of the path way association of the genes so that they will be able to know which gene will be up/down as a result of some particular gene being up/down.

References

1. Chad Creighton and Samir Hanash. (2003) Mining gene expression databases for association rules,vol 19 n0 1 pages: 79-86.
2. Patricia B. Cerrito. Combining SAS Text Miner with the Association Node in SAS Enterprise Miner™ to Investigate Inventory Data, SUGI 30 Proceedings, 076-30. 2005.
3. Xia, X. and Xie.Z. (2001) Amada: Analysis of microarray data. Bioinformatics 17:569-570.
4. Mussie Tesfamicael. (2004) Structural equation modeling assessing micro array data, Nashville, TN, SESUG.

CONTACT INFORMATION

Mussie Tesfamicael
Department of Mathematics
University of Louisville
Louisville, KY 40292
Work Phone: 502-852-7012, 502-298-8240
Fax: 502-852-7132
Email: matesf01@louisville.edu