

Graphical Modeling Based Gene Interaction Analysis for Microarray Data

Xintao Wu
UNC Charlotte
9201 University City Blvd.
Charlotte, NC 28269
xwu@uncc.edu

Yong Ye
UNC Charlotte
9201 University City Blvd.
Charlotte, NC 28269
yye@uncc.edu

Liyang Zhang
Memorial Sloan Kettering
Cancer Center
zhangl2@mskcc.org

ABSTRACT

DNA Microarray provides a powerful basis for analysis of gene expression. Data mining methods such as clustering have been widely applied to microarray data to link genes that show similar expression patterns. However, this approach usually fails to unveil gene-gene interactions in the same cluster. In this paper, we propose to use graphical modeling based interaction analysis for this purpose. We apply graphical gaussian model to discover pairwise gene interactions and use loglinear model to discover multi-gene interactions. We have constructed a prototype system that permits rapid interactive exploration of gene relationships; results can be validated by experts or known information, or suggest new experiments. We have tested our methodology using the yeast microarray data. Our results reveal some previously unknown interactions that have solid biological explanations.

Keywords

Graphical Modeling, Loglinear Modeling, Microarray Data Analysis

1. INTRODUCTION

With the completion of the human DNA sequence as part of the Human Genome Project [27], studies of gene-gene interactions are playing an increasingly important role in the search for the causes of human diseases. While individual genes may be responsible for making proteins, proteins usually interact in different physiological processes and pathways. Clues to the function of an unknown protein can be determined by investigating its interaction with other proteins whose functions are already characterized.

DNA microarrays provide a “snapshot” of transcription levels within the cell. It allows the simultaneous examination of thousands of genes in a single experiment. The raw microarray images are transformed into gene expression matrices where the rows usually denote genes and the columns denote various samples, conditions, or time points. The uniqueness of microarray data is that genes in rows are of very high dimensionality (e.g., $10^3 - 10^4$ genes) while samples in columns are of relatively low dimensionality (e.g., $10^1 - 10^2$ samples). A major challenge in computational biology is to uncover from such measurements, gene/protein interactions

and biological pathways at the molecular level. Exploration of coregulated genes can identify potential members of gene groups responsible for specific physiological processes.

In this paper, we investigate two graphical modeling techniques (Graphical Gaussian Modeling and Loglinear Modeling) for discovering gene interactions based on the correlation of their expression profiles. Graphical Gaussian Models (GGM) [28] assume a family of normal distributions for underlying data constrained to satisfy the pairwise conditional independence restrictions inherent in the independence graph. The independence graph generated from GGM is defined by a set of pairwise conditional independence relationships that determine the edge set of the graph. The weight of an edge denotes the partial correlation between two genes. The independence graph generated by graphical gaussian modeling can give domain users a basic understanding of interactions among a relatively large gene set. This large set might contain several pathways, as it is very likely that multiple signaling pathways interact with one another and the final biological response is shaped by interaction between pathways.

However, GGM can only detect dependencies that are close to linear. In particular, it is not likely to discover combinatorial effects (e.g., a gene has an extremely large probability to be over expressed only if several other genes are jointly over expressed, but not if at least one of them is not overexpressed). In this paper, we apply loglinear modeling [2], a methodology for approximating discrete multidimensional probability distributions, to discover the multi-way interactions. Loglinear modeling assumes multinomial distributions and needs a discretization of microarray expression data. For example, the gene expression values may be discretized to 2 categories, e.g., under-expressed and over-expressed, depending on whether the expression level is significantly lower than, or higher than control. The multi-way interactions have the potential to reveal complex (and often hidden) gene interactions, which cannot be discovered by other techniques (e.g., association rule, bayesian network, graphical gaussian model). However, the application of loglinear modeling is constrained by the size of samples as loglinear modeling requires the size of samples should be significantly larger than the number of cells in the contingency tables. For example, if the gene expression values are discretized to 2 categories, the contingency table built by 7 genes has 128 (2^7) cells which require more than 128 samples. In practice, many biological pathways and processes are known to involve interactions among a relatively large

number of genes. For instance, the human p53 signaling pathway contains 16 proteins, while the human integrin signaling pathway contains 36 proteins. Hence it is infeasible to apply loglinear modeling directly on those pathways with a relatively large number of genes. In this paper we also investigate graphical decomposition techniques to decompose independence graph into components and apply loglinear modeling on each component.

Our approach can effectively discover both pairwise interactions and multi-way interactions between genes. Furthermore, the edge weights of independence graph and the parameters of loglinear modeling can also be used to quantize the interactions between genes. We believe this work will complement current research on gene interactions [26; 4; 32] and can significantly contribute towards the biological annotations of genes including GENMAPP [9], Gene Ontology [3].

The remainder of the paper is structured as follows. In Section 2 we review the related work. Section 3 presents our interactive gene interaction analysis framework. We present details of graphical modeling in Section 4. Experimental results are discussed in Section 5. In Section 6 we draw conclusions and describe directions for future work.

2. RELATED WORK

Clustering algorithms (e.g., CAST [5], MST [33], HCS [15], CLICK [22], BICLUSTER [7]) have been quite successful in the molecular profiling of human cancers, however they are insufficient to identify molecular networks. It is impossible to determine the interactions that can exist between different genes from one cluster, especially when a gene can participate in more than one gene network.

The graphical gaussian models were previously applied for gene expression analysis in [18] where they applied multiple regression procedures with variable selection to get approximate partial correlations between any pair of genes. However, multiple regression procedures are infeasible for microarray data sets with thousands of genes because of high computational cost.

In [8], association rules [1] are applied to investigate how the expression of one gene may be associated with the expression of a set of genes. The kind of rule can be discovered is, for example, “when gene A and gene B are over expressed within a sample, then often gene C is also over expressed”. Theoretically, the association rule method is able to resolve the drawbacks of existing clustering approaches by assigning a gene to many subsets, however, the association rule method can only capture gene co-expression, and not interactions because it is exclusively based on support measure. Some measures, such as lift [23], pairwise associations [10] have been investigated to overcome the limitations of support-based association algorithms. In this paper, we extend and generalize the previous work by the all k -way interaction model. k -way relationships have the potential to reveal complex (and often hidden) gene interactions, which cannot be explained by any low level interactions. Furthermore, our model can also interpret the interestingness of associations by examining loglinear parameters.

Both graphical gaussian modeling and loglinear modeling are based on correlation measure instead of causality measure. Bayesian network, which is based on directed acyclic graph (DAG) and can provide models of causal influence,

has recently been investigated for gene regulatory networks [13; 21]. The advantage of bayesian network is that it generates a directed graph that suggests causal influence. However, bayesian network cannot discover multi-way effects as it assumes only linear interactions. Another difficulty with this technique is that learning the bayesian network structure is an NP-hard problem, as the number of DAGs is superexponential in the number of genes, and exhaustive search is intractable.

3. THE FRAMEWORK OF INTERACTIVE ANALYSIS OF GENE INTERACTIONS

In microarray data sets, genes in rows and samples in columns are of very different dimensionality (e.g., $10^3 - 10^4$ genes versus $10^1 - 10^2$ samples). Thus, microarray data sets are very sparse in high-dimensional gene spaces. Our approach is to explore inter-relationships between a subset of genes. Figure 1 shows the framework of our prototype system for interactive gene interaction analysis. Specifically, it involves the following steps:

1. Preprocessing: We subject the input data to hierarchical clustering or association rule mining, prior to analyzing gene interactions.
2. Graphical Gaussian Modeling: Subsets of genes (clusters or frequent itemsets) are then analyzed for pairwise gene interaction using GGMs.
3. Decomposition: The independence graph from graphical gaussian models is then decomposed to get components.
4. Loglinear Modeling: The genes included in each component are then analyzed to get multi-way effects by using loglinear models.
5. Visualization: The user may explore the output of both GGMs and loglinear models interactively.

The reason we subject the input data to hierarchical clustering or association rule mining prior to analyzing gene interactions has two folds. First, due to the large data size, it is infeasible to apply the GGMs directly to the original data; Second, the correlation matrix is inevitably degenerate, as the matrix rank is bounded by the sample size. Here in our framework, the number of genes contained in each cluster or frequent itemset is usually less than the size of samples, which avoids the matrix rank problem.

The graphical gaussian modeling method is statistically sound and computationally tractable for analyzing microarray data and inferring pairwise biological interactions from them. As the number of cells in contingency table (which is determined by the number of genes and the number of categories for each gene) may significantly exceed the number of samples, it may be inaccurate to apply for loglinear modeling directly on each subset. Hence, we decompose each subset into components by graphical decomposition techniques and each component is then analyzed by loglinear models.

Given the inaccuracies and limitations of clustering and association rule mining, one cannot assume that the identified subsets of genes are completely independent of the remaining genes of the whole genome. Thus, we apply interactive techniques whereby a user can interactively analyze gene

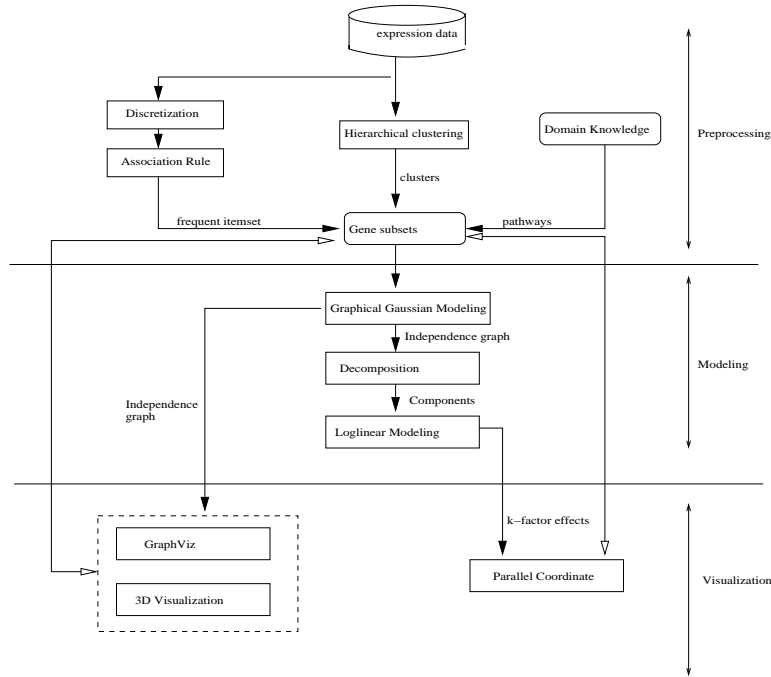


Figure 1: The framework of prototype system of gene interaction analysis

interactions by adding or removing any number of genes to/from one subset. To make this interactive exploration intuitive and efficient, we apply information visualization techniques, whereby visual representations present the interface to interactive exploration. In this work, we use automatic graph drawing techniques to display and edit gene subsets and their 2-way relationships and use parallel coordinate techniques to display multi-gene interactions from loglinear models. We are also working on interactive visual representations for cluster hierarchies as well as association rule mine sets, so as to rapidly focus, view and interactively edit gene subsets of interest.

4. GRAPHICAL INTERACTION ANALYSIS METHODS

Let $\mathcal{S} = \{s_1, s_2, \dots, s_m\}$ be the set of samples or conditions and $\mathcal{G} = \{g_1, g_2, \dots, g_n\}$ be the set of genes. The microarray data can be represented as $\mathcal{X} = \{x_{ij} \mid i = 1, \dots, n, j = 1, \dots, m\}$ ($n \gg m$), where x_{ij} corresponds to the expression value of the sample s_j on gene g_i .

In Section 4.1 and Section 4.2, we discuss graphical gaussian modeling and loglinear modeling respectively. We leave the discussion of graphical decomposition in Section 4.3 and assume the number of genes is low or medium (depending on the number of samples and the discretization levels for each gene).

4.1 Graphical Gaussian Modeling

Graphical Gaussian Models, also known as covariance selection models, assume a family of normal distributions for underlying data constrained to satisfy the pairwise conditional independence restrictions inherent in the independence graph. The microarray expression data, which are log

transformed from the raw microarray images, satisfy near multivariate normal distribution due to the nature of experimental errors. The independence graph is defined by a set of pairwise conditional independence relationships that determine the edge set of the graph. A crucial concept of applying GGM is that of partial correlation. That is, measuring the correlation between two variables after the common effects of all other variables in the genome are removed.

$$pr_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}} \quad (1)$$

Equation 1 shows the form for partial correlation of two genes g_x and g_y while controlling for a third gene variable g_z , where r_{xy} denotes Pearson's correlation coefficient. The partial correlation ($pr_{xy.z}$) of genes g_x and g_y with respect to gene g_z may be considered to be the correlation (r_{xy}) of g_x and g_y after the effect of g_z is removed. If there is no difference between $pr_{xy.z}$ and r_{xy} , we can infer that the control variable g_z has no effect. If the partial correlation approaches zero, the inference is that the original correlation is spurious (i.e., there is no direct causal link between the two original gene variables because the control gene variable is either common antecedent cause, or intervening variables). Partial correlations that remain significantly different from zero may be taken as indicators of a possible causal link.

It is important to note that partial correlation is different from standard correlation, indicates better evidence for genetic regulatory links than standard correlation, as shown from our preliminary results [31], and is in agreement with biological interpretation. Consider the example in Figure 2, that shows the correlation and partial correlation graph over a subset of genes from the yeast data[16]. Figure 2(b) shows pairwise correlations with correlation coefficient greater than 0.65, which indicates positive correlations between any pair

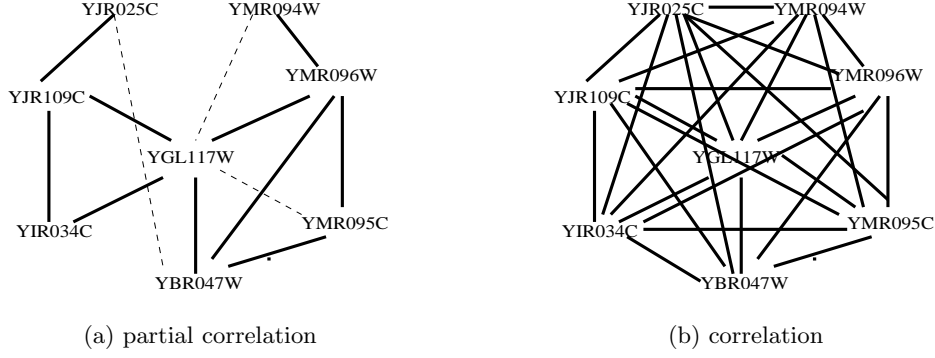


Figure 2: Gene interactions using partial correlation vs. correlation, the threshold for partial correlation is 0.2 while the threshold for correlation is 0.65. Note dashed lines indicate a negative partial correlation and solid lines indicate a positive partial correlation.

of genes. However, partial correlations in Figure 2(a) show no interaction between 15 pairs of genes (genes with high correlations may be controlled by a common gene and not directly linked in the pathway) and even negative interactions between three pairs of genes.

With a set of genes g , the partial correlation can be computed by $pr_{xy.g} = -\frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$, where s_{xy} is the xy -th element of the inverse of variance matrix ($\mathcal{S} = \mathcal{V}^{-1}$). It is known that conditional independence constraints are equivalent to specifying zeros in the inverse variance [28]. The method can be sketched as follows:

- Compute the variance matrix \mathcal{V} where v_{ij} , $i, j = 1, \dots, n$, corresponds to covariance between gene g_i and g_j .
- Compute its inverse $\mathcal{S} = \mathcal{V}^{-1}$.
- Scale \mathcal{S} to have a unit diagonal and compute partial correlations $pr_{x_i x_j . g}$.
- Draw the independence graph according to the rule that no edge is included in the graph if the absolute value of partial correlation coefficient is less than some threshold.
- Fit GGMs by maximum likelihood estimation.

4.2 Loglinear Modeling

In this section we describe in detail how we screen gene interactions by means of building all k -way interaction models and examining their parameters and residuals using microarray data. Here we assume a set of components, $\mathcal{S}^{(0)}$, are given by graphical decomposition techniques, users, or maximal frequent itemset by using Apriori algorithm. For each component, we build all k -way interaction models iteratively, and screen large gene sets based on the estimates from k -way interaction model. The method can be sketched as follows:

- Discover to get component set $\mathcal{S}^{(0)}$
- For $k=1$ to K
 - For each large gene set $s \in \mathcal{S}^{(k-1)}$

- fit k -way interaction model
- if its standardized residual $e^{(k)} > \tau$
- include s into $\mathcal{S}^{(k)}$

The key to finding interactions worthy of examining (those that will join the lists $\mathcal{S}^{(k)}$), is to compute its standardized residual $e^{(k)}$. Equation 2 shows the standardized residual form used in our framework, where y is the actual support of s and $\hat{y}^{(k)}$ is the estimated value given by the k -way interaction model.

$$e^{(k)} = \frac{y - \hat{y}^{(k)}}{\sqrt{\hat{y}^{(k)}}} \quad (2)$$

When the model holds, $e^{(k)}$ is asymptotically normal with mean 0. In comparing standardized residuals to standard normal percentage points, we obtain conservative indications of cells having lack of fit. When the residual is large, it means that the support of s cannot be explained by the k -way interactions, thus higher order interactions (larger than k) are at play.

4.2.1 Loglinear Model Revisited

Loglinear modeling is a methodology for approximating discrete multidimensional probability distributions. The multi-way table of joint probabilities is approximated by a product of lower-order tables.

Given a value $y_{i_1 i_2 \dots i_n}$ at position i_r of the r th dimension d_r ($1 \leq r \leq n$), we define the log of anticipated value $\hat{y}_{i_1 i_2 \dots i_n}$ as a linear additive function of contributions from various higher level group-bys as

$$\hat{l}_{i_1 i_2 \dots i_n} = \log \hat{y}_{i_1 i_2 \dots i_n} = \sum_{G \subseteq \{d_1, d_2, \dots, d_n\}} \gamma_{(i_r | d_r \in G)}^G \quad (3)$$

where the γ terms are the coefficients of the model. The coefficients corresponding to any group-by G are obtained by subtracting from the average l value at group-by G all the coefficients from higher level group-by-s.

For instance, in a 4-dimensional table with dimensions A, B, C, D , we use (i, j, k, l, y_{ijkl}) to denote the cell in a 4-D cube space,

where $i = 0, \dots, I-1, j = 0, \dots, J-1, k = 0, \dots, K-1, l = 0, \dots, L-1$. Equation 4 shows the saturated loglinear model which contains all the possible k -factor effects, all the possible $k-1$ -factor effects, and so on up to the 1-factor effects and the mean γ .

$$\begin{aligned} \log \hat{y}_{ijkl} = & \gamma + \gamma_i^A + \gamma_j^B + \gamma_k^C + \gamma_l^D \\ & + \gamma_{ij}^{AB} + \gamma_{ik}^{AC} + \gamma_{il}^{AD} + \gamma_{jk}^{BC} + \gamma_{jl}^{BD} + \gamma_{kl}^{CD} \\ & + \gamma_{ijk}^{ABC} + \gamma_{ijl}^{ABD} + \gamma_{ikl}^{ACD} + \gamma_{jkl}^{BCD} \\ & + \gamma_{ijkl}^{ABCD} \end{aligned} \quad (4)$$

For example, γ_i^A is one-factor effect, γ_{ij}^{AB} is two-factor effect which shows the dependency within the distributions of the associated attributes A and B , γ_{ijk}^{ABC} is three-factor effect which shows the dependency within the distributions of all the associated attributes A, B , and C . It is important to note the multiple-factor effects can capture the complex interactions such as catalysis and cooperativity in biology. For example, if all two-factor effects of A, B, C ($\gamma_{ij}^{AB}, \gamma_{ik}^{AC}, \gamma_{jk}^{BC}$) are insignificant (close to 0) and the three-factor effect (γ_{ijk}^{ABC}) may well describe the rule such as “a gene is over (or under) expressed only if several genes are jointly over (or under) expressed”.

The loglinear theory requires the loglinear parameters sum to 0 over all indices. For example, $\gamma_i^A = \sum_{j=0}^{J-1} \gamma_{ij}^{AB}$, where a dot “.” means that the parameter has been summed over the index. Equation 5 shows how to compute the coefficients in a 4-dimensional table.

$$\begin{aligned} \gamma &= l_{\dots} \\ \gamma_i^A &= l_{i\dots} - \gamma \\ &\dots \\ \gamma_{ij}^{AB} &= l_{ij..} - \gamma_i^A - \gamma_j^B - \gamma \\ &\dots \end{aligned} \quad (5)$$

In [20] a fast computation technique called the UpDown method that makes this approach feasible for large sets is described. In this paper we apply UpDown approach to compute the parameters of all k -way interaction models.

4.2.2 All k -way Interaction Loglinear Model Fitting

For each component, we build one contingency table which will be analyzed by all k -way loglinear models. Table 1 shows one contingency table built from yeast data based on one component with four genes (e.g., YHR071W, YMR094W, YMR096W, YMR095C). Each expression is discretized into three categories: overexpressed, normal, and underexpressed. It is important to notice that gene expression values can be discretized into any number of categories and our k -way interaction loglinear model can be built directly over the transformed contingency table.

$$\log \hat{y}_{ijkl}^{(1)} = \gamma + \gamma_i^A + \gamma_j^B + \gamma_k^C + \gamma_l^D \quad (6)$$

$$\begin{aligned} \log \hat{y}_{ijkl}^{(2)} = & \gamma + \gamma_i^A + \gamma_j^B + \gamma_k^C + \gamma_l^D + \gamma_{ij}^{AB} \\ & + \gamma_{ik}^{AC} + \gamma_{il}^{AD} + \gamma_{jk}^{BC} + \gamma_{jl}^{BD} + \gamma_{kl}^{CD} \end{aligned} \quad (7)$$

$$\begin{aligned} \log \hat{y}_{ijkl}^{(3)} = & \gamma + \gamma_i^A + \gamma_j^B + \gamma_k^C + \gamma_l^D + \gamma_{ij}^{AB} \\ & + \gamma_{ik}^{AC} + \gamma_{il}^{AD} + \gamma_{jk}^{BC} + \gamma_{jl}^{BD} + \gamma_{kl}^{CD} \\ & + \gamma_{ijk}^{ABC} + \gamma_{ijl}^{ABD} + \gamma_{ikl}^{ACD} + \gamma_{jkl}^{BCD} \end{aligned} \quad (8)$$

Equation 6 and 7 shows all 1-way and all 2-way interaction model respectively. Equation 6 assumes the independence model and includes all-one-factor (main) effects and grand mean. Equation 7 includes all-two-factor effects apart from all-one-factor effects and grand mean. The comparison between the observed value y with either $\hat{y}^{(1)}$ or $\hat{y}^{(2)}$ is used to screen interesting item sets in [23] or [10] respectively. However, the assumed independence model or pairwise model may be insufficient to fit some high factor gene interactions. In [10], they only distinguish between multi-item associations that can be explained by all pairwise associations, and item sets that are significantly more frequent than their pairwise associations would suggest. In our framework, we extend to all k -way interaction models (e.g., as shown in Equation 8). Furthermore, we may interpret associations by examining the γ -terms of fitted loglinear models instead of by only examining the differences between observed frequencies of item sets and expected frequencies computed from assumed models.

4.2.3 Interpreting Interactions by Examining Parameters

If the gene expression values are discretized to 2 categories, over-expressed and under-expressed, our previous results [29] have two important conclusions:

- Each of the γ -term has only one absolute value due to linear constraints of coefficients and the positive (negative) value implies positive (negative) associations.
- We can compare the interactions according to their magnitude of γ -terms derived from loglinear models.

Figure 3 shows the parameter values from the saturated model. Each of the γ -term in the saturated loglinear model describes the interaction of item variables. For example, γ^{AB} represents the interaction between gene A and B. For example, $\gamma^{AB} = 0.275$ in Figure 3 implies $\gamma_{00}^{AB} = 0.275$, $\gamma_{01}^{AB} = -0.275$, $\gamma_{10}^{AB} = -0.275$, and $\gamma_{11}^{AB} = 0.275$. It can be interpreted that the overexpression (underexpression) of A implies the overexpression (underexpression) of B with interaction effect of 0.275. Furthermore, the comparison of γ^{BD} (0.278) and γ^{AD} (0.052) implies the interaction of BD is more significant than that of AD.

Though two-category discretization is enough for most cases (especially during exploratory phase), the users may need to investigate the interactions at finer levels (e.g., what is the effect of weak-overexpressed of gene A on gene B) which requires multiple-category discretization. In this case we cannot compare the magnitude of γ -terms directly. This is due to several reasons. Firstly, the degree of freedom (d.f.) for each particular interaction varies (however, in two-category case, the d.f. for each particular interaction is always 1).

Table 1: One contingency table built from yeast data with four genes where A,B,C,D denotes YHR071W, YMR094W, YMR096W, YMR095C respectively. The cell $(A \uparrow, B \uparrow, C \uparrow, D \uparrow)$ with value 54 is large item set discovered by association rule method.

		$B \downarrow$			B			$B \uparrow$		
		$A \downarrow$	A	$A \uparrow$	$A \downarrow$	A	$A \uparrow$	$A \downarrow$	A	$A \uparrow$
$D \downarrow$	$C \downarrow$	5	5	0	4	9	0	0	0	0
		C	0	0	0	0	7	0	0	0
		$C \uparrow$	0	0	0	0	0	0	0	0
D	$C \downarrow$	1	0	0	3	7	0	0	0	0
		C	0	0	0	4	130	7	0	1
		$C \uparrow$	0	0	1	0	7	2	0	1
$D \uparrow$	$C \downarrow$	0	0	0	0	0	0	0	0	0
		C	0	0	0	1	11	0	0	0
		$C \uparrow$	0	0	0	0	15	3	0	19
										54

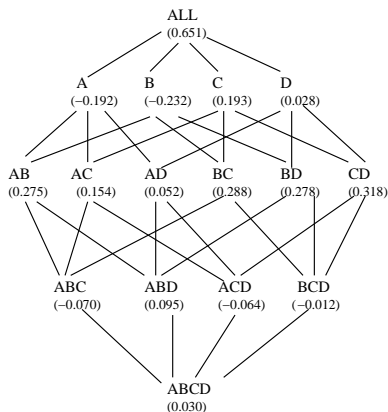


Figure 3: Lattice for the data set with four dimensions denoted by A, B, C, D respectively. The value in $()$ denotes the value of γ -term of saturated loglinear model

Secondly, the variance for each interaction varies (in two-category case, the variances for all γ -terms are the same). So in the general case, we have to compute the standardized parameter value $(\gamma/\sigma(\gamma))$ for each γ -term in order to compare the significance of each interaction. Thirdly, there can be more than one absolute value for each γ -term and we have to combine the estimates in some way to form an overall test statistic [14].

All k-way interaction loglinear models are built from transformed contingency table where we may lose some information due to discretization when preprocessing raw data. The all 2-way interaction loglinear model is a direct parallel to the graphical gaussian model. In both the all 2-way interaction loglinear model and the graphical gaussian model, conditional independence between any pair of genes is parameterised by a single scalar, the mixed derivative measure of partial interaction. We can see there is no information loss in graphical gaussian models as we do not need to discretize the expression values. However, the graphical gaussian models can not capture k-way ($k > 2$) interactions.

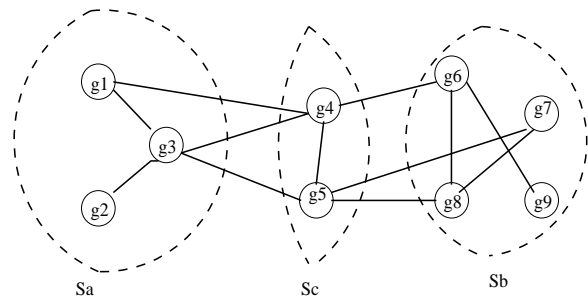


Figure 4: Graphical decomposition

4.3 Graphical Decomposition

As we stated in the introduction, we cannot build loglinear models over the very sparse and large contingency table that results from gene subsets with large number of genes. For example, Figure 4 shows an independence graph with 9 genes. If each expression value is discretized into three categories, the number of cells in the contingency table will be 3^9 which is much larger than samples (usually $10^1 - 10^2$). Hence we need to decompose independence graph into subgraphs.

Graph-theoretical results show that if a graph corresponding to a graphical model is decomposable into subgraphs by a clique separator¹, the MLEs for the parameters of the model can easily be derived by combining the estimates of the models on the simpler subgraphs. Hence, applying a divide-and-conquer approach based on the decompositions will make the procedure applicable to much larger subsets of genes.

The theory may be interpreted by the following way as shown in Figure 4: if two disjoint subsets of vertices S_a and S_b are separated by a subset S_c in the sense that all paths from S_a to S_b go through S_c , then the variables in S_a are conditionally independent of those in S_b given the vari-

¹A clique is a subset of vertices which induce a complete subgraph for which the addition of any further vertex renders the induced subgraph incomplete. A graph is complete if all vertices are joined with undirected edges. In other words, the clique is maximally complete.

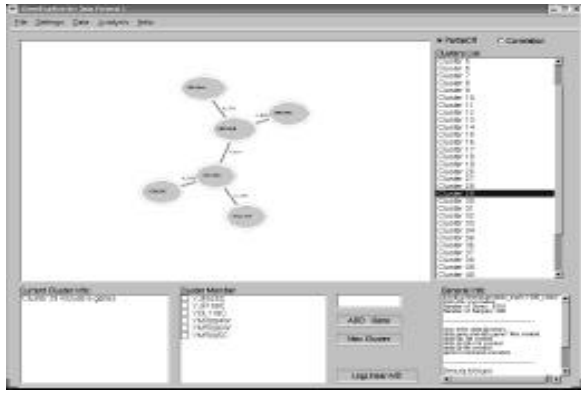


Figure 5: Snapshot of prototype system for gene interaction analysis

ables in S_c . The subgraphs may be further decomposed into subgraphs $S_a \cup S_c$ and $S_b \cup S_c$. The requirement that the subgraph on S_c is complete implies that there is no further independence constraints on the elements of S_c , so that this factorization contains all the information about the joint distribution. To find the clique separators of a graph or to find the vertex-sets of the irreducible components of the graphs, an algorithm with a complexity of $O(ne + n^2)$ can be used [25], where n is the number of vertices and e is the number of edges.

5. EXPERIMENTAL RESULTS

The experiments were conducted in a DELL Precision 340 Workstation (Redhat Linux 9.0 operating system), with one 2.4G processor, and 1G bytes of RAM.

In this section we show the results on yeast data [16] which contains expression profiles for 6316 transcripts corresponding to 300 diverse mutations and chemical treatments in yeast. In [8], this yeast data set is transformed by binning 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. We apply the same discretization strategy in our experiment.

Figure 5 shows a snapshot of our prototype system for gene interaction analysis. Our system has features that allow its users to choose frequent itemset mining [6] and various clustering methods [11] to get subsets of genes. For each subset, the independence graph is generated by using GGMs. The users may interactively add or remove some genes from the independence graph and the new independence graph will be generated interactively. We use automatic graph drawing tools [12] to represent gene networks. Our implementation is in C++ on Unix workstations using FLTK [24] for the user interface.

5.1 Preprocessing

The preprocessing data mining techniques to get gene subsets we applied in this experiment include frequent item set and maximal frequent item set mining method with different support, K-mean, SOM, PCA, and Hierarchical clustering methods. Table 2 shows the size of gene sets obtained using frequent itemset and maximal frequent itemset min-

Table 2: Size of gene sets obtained using frequent itemset and maximal frequent itemset with different support and execution time (in seconds) of frequent itemset mining (T_f), GGMs (T_g) and loglinear modeling (T_l)

support(%)	frequent	max. frequent	T_f	T_g	T_l
12	130603	1635	1.25	0.03	1703
13	22123	795	0.35	0.02	905
14	2735	298	0.12	0.01	291
15	1134	164	0.11	0.01	172
16	314	69	0.10	0	3.5
17	79	23	0.10	0	1.2
18	39	16	0.10	0	0.23
19	17	10	0.10	0	0.15
20	8	4	0.09	0	0.15

Table 3: Execution time (T_c) using various clustering methods

clustering	execution time T_c (s)
K-mean	1205
SOM	1382
PCA	1021
Hierarchical	7205

ing method with different support. We can see the size of frequent item set and maximal frequent item set under low support values is large.

Table 2 also shows the execution time of preprocessing (frequent itemset mining by Apriori) and that of GGMs and loglinear modeling over all subsets. We can see the execution time of GGMs is trivial as it is even less than that of frequent itemset mining. However, the execution time of loglinear modeling is increasing significantly when more subsets need to be built on. As loglinear modeling is mainly used for interactive analysis, the execution for each subset is still fast. Table 3 shows the execution times using different clustering methods which are larger than that of loglinear modeling.

5.2 Pairwise Interaction using GGMs

Figure 6 demonstrates the pairwise interaction for one selected gene group with 11 genes (We omit biological information for each gene due to space limitation.). Briefly, nine genes have known functions and seven genes encode proteins that are involved in biosynthesis/metabolism. Some facts that can be inferred from the interaction graph include:

- There are two groups of genes with known functions where the partial correlation between genes within each group is greater than 0.3. They are: 1) YMR095C - YMR096W - YMR094W 2) YJR109C - YIL116W - YIR034C - YDL198C - YJR109C. This indicates the expression of those genes are highly correlated, which agrees with laboratory data.
- YMR029C is not connected with any other genes. As

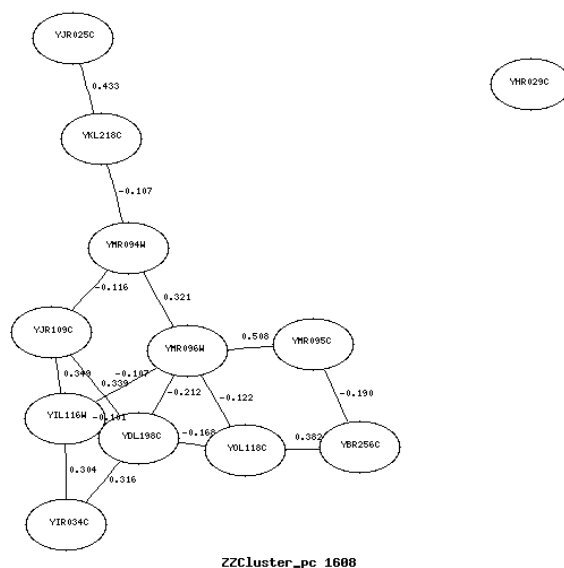


Figure 6: Pairwise gene interactions using GGMs for a selected maximal frequent gene set

this gene has no correlation with the remaining genes, we may remove this gene from gene subsets though this gene is included in the frequent itemset from association rule mining.

- The negative correlation (e.g., between YMR095C and YBR250C) in Figure 6 indicates that the functions of each pair of genes may counteract with each other (activators and repressors) of the biosynthesis/metabolism pathways or their expression is negatively regulated by the other gene in each pair.

Our results receive some solid biological explanations. For example, SNZ1 (YMR096W) belongs to three-member gene families SNZ1-3 whereas SNO1 (YMR095C) belongs to another three-membered gene families SNO1-3 (Snz-proximal open reading frame). The DNA sequences and relative positions of SNZ and SNO genes have been phylogenetically conserved. SNZ-SNO gene pairs are co-regulated under various conditions. Recent studies indicated that SNZ1 and SNO1 are involved in cellular responses to nutrient limitation. Both of them are required for yeast to grow in pyridoxine (vitamin B6) lacking media, indicating that they are involved in pyridoxine metabolism.

5.3 Multi-way Interaction using Loglinear

Table 4 shows size of gene sets using k-way interaction model with different support. We can see many gene sets are screened by all k-way interaction model when we increase k .

Table 5 shows the frequencies and estimates from all k-way interaction model for Large 4-gene sets². For the component with four genes YJR109C, YMR094W, YMR096W, and YMR095C (line 3 in table 5, we use A, B, C, D to

²The information of each ORF (open reading frame) can be retrieved from the Saccharomyces Genome Database (<http://genome-www.stanford.edu/Saccharomyces/>).

denote each gene respectively), we get multi-way interactions among ABC (-0.006), ABD (0.108), ACD (-0.01), and BCD (0.127). All the pairwise interactions are above 0.13. We can easily derive significant positive interactions exist among ABD and BCD while non-significant negative interactions exist among ABC and ACD.

We can see our results agree to some previously known biological interactions or reveal some previously unknown interactions that have solid biological explanations.

- YMR096W (SNZ1) and YMR095C (SNO1) are present in 8/8 groups in Table 5, while YMR096W (SNZ1), YMR095C (SNO1) and YMR094W (CTF13) is present in 6/8 groups in Table 5. The DNA sequences and relative positions of SNZ and SNO genes have been phylogenetically conserved. SNZ-SNO gene pairs are coregulated under various conditions [19].
- CTF13, SNO1 and SNZ1, located adjacent to each other, are situated proximal to the centromere on the right arm of chromosome XIII. We project that the co-regulation of these three genes might be caused by the conformational changes of chromosomal structure during transcription activation even though the possibility that they are involved in the same biological process is not excluded.
- YJR109C (CPA2) encodes one of the two subunits of carbamoylphosphate synthase in the arginine synthesis pathway. The expression of CPA2 is increased when arginine is limited [17]. The overexpression of CPA2 indicated that certain conditions in Hughes experiments may somehow limited arginine which leads to increased expression of CPA2. The co-regulation of CPA2 and SNO1/SNZ1 implies that they might be involved in the same biological process.

6. CONCLUSIONS

Table 4: Size of gene sets obtained using k-way interaction model with different support. $S^{(k)}$ ($k = 0, 1, 2, 3$) denotes the size of item sets which can not be interpreted by all k-way interaction models.

support(%)	$\ S^{(0)}\ $	$\ S^{(1)}\ $	$\ S^{(2)}\ $	$\ S^{(3)}\ $
14	2735	2500	2253	1931
15	1134	1084	852	691
18	39	39	19	8
20	8	8	4	1

Table 5: The frequencies and estimates from all k-way interaction model for Large 4-gene components. All of the genes listed in each set represent the gene being up in the sample.

Gene Set	Frequency	1-way	2-way	3-way
YHR029C, YMR094W, YMR096W, YMR095C	56	0	15	26
YJR109C, YGL117W, YMR096W, YMR095C	54	0	15	23
YJR109C, YMR094W, YMR096W, YMR095C	56	0	17	32
YGL117W, YER175C, YMR096W, YMR095C	54	0	24	28
YGL117W, YMR094W, YMR096W, YMR095C	56	0	21	27
YHR071W, YMR094W, YMR096W, YMR095C	54	0	22	33
YBR047W, YMR094W, YMR096W, YMR095C	59	0	14	18
YER175C, YMR094W, YMR096W, YMR095C	61	0	20	24

In this paper we have applied a combination of graphical gaussian modeling and loglinear modeling to find meaningful interactions among sets of genes in gene expression data collected by microarrays. The graphical gaussian modeling can effectively discover pairwise interactions between genes while the loglinear modeling can discover multi-way interactions hidden in components. We have shown that the application of the method to yeast microarray data uncovers a set of interactions that can be explained using biological arguments, and thus are meaningful. As such, we believe that this method complements the typical clustering approaches used to analyze microarray data.

There are some aspects of this work that merits further research. Among them, we are studying how to better deal with sparse data when either structural zero cells present or it contains many small cell values. It is known that loglinear model can still work for small incomplete table with structural or sampling zeros. We will investigate other techniques such as shrinkage estimates [10] for large sparse microarray data.

7. ACKNOWLEDGEMENTS

This paper is an extension and combination of two previous workshop papers [31; 30] and was supported, in part, by funds provided by the University of North Carolina at Charlotte. The authors would like to thank Christian Borgelt for his implementation of the Apriori algorithm and Michael Ellison for his implementation of clustering program which makes our system possible. Our system can be downloaded via <http://www.cs.uncc.edu/xwu/bio/GenExplore.html>. We also would like to thank Daniel Baraba and Kalpathi Subramanian for useful discussions in building the prototype system.

8. REFERENCES

- [1] R. Agrawal, T. Imilienski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Database*, pages 207–216, 1993.
- [2] E. Andersen. *The statistical analysis of categorical data*. Springer Verlag, Berlin, Heidelberg, 1994.
- [3] M. Ashburner, C. Ball, J. Blake, and et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25:25–29, 2000.
- [4] G. Bader and C. Hogue. Bind: a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics*, 16:465–477, 2000.
- [5] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3/4):281–297, 1999.
- [6] C. Borgelt. Association rule induction. <http://fuzzy.cs.uni-magdeburg.de/borgelt/software.html>.
- [7] Y. Cheng and G. Church. Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 93–103. San Diego, CA, August 2000.
- [8] C. Creighton and S. Hanash. Mining gene expression databases for association rules. *Bioinformatics*, 19:1:79–86, 2003.

- [9] K. Dahlquist, N. Salomonist, K. Vranizan, S. Lawlor, and B. Conklin. Genmapp, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet*, 31:19–20, 2002.
- [10] W. DuMouchel and D. Pregibon. Empirical bayes screening for multi-item association. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data*. San Francisco, CA, August 2001.
- [11] M. Eisen. Cluster analysis and visualization. <http://rana.lbl.gov>.
- [12] J. Ellson and S. North. Graph visualization project (graphviz). <http://www.graphviz.org>.
- [13] N. Friedman, M. Linial, I. Nachman, and D. Peer. Using bayesian networks to analyze expression data. In *Proceedings of the fourth Annual International Conference on Computational Molecular Biology*, 2000.
- [14] L. Goodman. The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics*, 13:33–61, 1971.
- [15] E. Hartuv and R. Shamir. A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76(4-6):175–181, 2000.
- [16] T. Hughes, M. Marton, A. R. Jones, C. Roberts, R. Stoughton, C. Armour, H. Bennett, E. Coffey, H. Dai, Y. He, M. J. Kidd, and A. M. King. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.
- [17] D. Kinney and C. Lusty. Arginine restriction induced by delta-n-(phosphonacetyl)-l-ornithine signals increased expression of his3, trp5, cpa1, and cpa2 in saccharomyces cerevisiae. *Mol. Cell Boil*, 9:4882–4888, 1989.
- [18] H. Kishino and P. J. Waddell. Correspondence analysis of genes and tissue types and finding genetic links from microarray data. *Genome Informatics*, 11:83–95, 2000.
- [19] P. A. Padilla, E. K. Fuge, M. E. Crawford, A. Errett, and M. Werner-Washburne. The highly conserved, coregulated sno and snz gene families in saccharomyces cerevisiae respond to nutrient limitation. *Journal of Bacteriol*, 180:5718–5726, 1998.
- [20] S. Sarawagi, R. Agrawal, and N. Meggido. Discovery-driven exploration of olap data cubes. In *Proceedings of the International Conference on Extending Data Base Technology*, pages 168–182. Valencia, Spain, 1998.
- [21] E. Segal, M. Shapira, A. Regev, D. Peer, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34:166–176, 2003.
- [22] R. Shamir and R. Shamir. Click: A clustering algorithm for gene expression analysis. In *Proceedings of the Eighth International Conference on Intelligent System for Molecular Biology (ISMB00)*, 2000.
- [23] C. Silverstein, S. Brin, and R. Motwani. Beyond market baskets: generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 2:39–68, 1998.
- [24] B. Spitzak and et. al. The fast light toolkit(fttk). <http://www.fttk.org>.
- [25] R. Tarjan. Decomposition by clique separators. *Discrete Mathematics*, 55:221–232, 1985.
- [26] P. Uetz, L. Giot, G. Cagney, and et al. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403:623–627, 2000.
- [27] J. Venter, M. Adams, E. Myers, and et al. The sequence of the human genome. *Science*, 291:1304–1351, 2001.
- [28] J. Whittaker. *Graphical Models in Applied Mathematical Multivariate Statistics*. Wiley, 1990.
- [29] X. Wu, D. Barbará, and Y. Ye. Screening and interpreting multi-item associations based on log-linear modeling. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 276–285. Washington, DC, August 2003.
- [30] X. Wu, D. Barbará, L. Zhang, and Y. Ye. Gene interaction analysis using k-way interaction loglinear model: A case study on yeast data. In *ICML03 Workshop on Machine Learning in Bioinformatics*, pages 38–45. Washington, DC, August 2003.
- [31] X. Wu, Y. Ye, K. Subramanian, and L. Zhang. Interactive gene interaction analysis using graphical gaussian models. In *The 3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics*, pages 63–69. Washington, DC, August 2003.
- [32] I. Xenarios, L. Salwinski, X. Duan, P. Higney, S. M. Kim, and D. Eisenberg. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 30:303–305, 2002.
- [33] Y. Xu, V. Olman, and D. Xu. Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics*, 18(4):536–545, 2002.