
Gene Interaction Analysis Using k-way Interaction Loglinear Model: A Case Study on Yeast Data

Xintao Wu

XWU@UNCC.EDU

University of North Carolina at Charlotte, CS Dept., 9201 University City Blvd., Charlotte, NC 28223

Daniel Barbará

DBARBARA@GMU.EDU

George Mason University, ISE Dept., Fairfax, VA 22303

Liyang Zhang

ZHANGL2@MSKCC.ORG

Memorial Sloan Kettering Cancer Center, New York, NY 10021

Yong Ye

YYE@UNCC.EDU

University of North Carolina at Charlotte, CS Dept., 9201 University City Blvd., Charlotte, NC 28223

Abstract

Microarray data 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 multiple interactions by the same gene. Association rule mining has been used for this purpose, but the inherent limitations of association rules limit the applicability of the results. In this paper we use a combination of association rule mining and loglinear modeling to discover k-gene interactions. Using this technique we can discover interactions among k-genes that cannot be explained by the combined effects of any of the subsets of those genes. We test our technique experimentally, using yeast microarray data. Our results reveal some previously unknown associations that have solid biological explanations.

1. Introduction

With the description of complete genome sequences, DNA microarray technology has become a powerful means for genome-wide expression profiling and analysis. 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). Hence such data sets are very sparse in high dimensional genes space. Furthermore, some of the genes collected may not necessarily be of interest or relevant. The challenge now is to rapidly and efficiently extract useful information and discover knowledge from such huge data sets such as gene functions, gene interactions, regulatory pathways, metabolic pathways, and effects of environmental factors.

Clustering algorithms (e.g., CAST (Ben-Dor et al., 1999), MST (Xu et al., 2002), HCS (Hartuv & Shamir, 2000), CLICK (Shamir & Shamir, 2000)) have been quite successful in the molecular profiling of human cancers, however they are insufficient to identify molecular networks. The goal of most clustering methods is to define each gene as being part of a self-contained cluster. Hence, each gene is assigned to only one cluster (in clustering, a gene cannot belong to two unrelated clusters in the hierarchy of clusters). However, a gene can usually be characterized in more than one way (the p53 protein belongs more than one physiological pathways). Furthermore, 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 authors, in (Creighton & Hanash, 2003), apply

association rules (Agrawal et al., 1993) to investigate how the expression of one gene may be associated with the expression of a set of genes. One might infer that genes involved participate in some type of gene network based on association rules or frequent item sets. Association rules are defined by the support of the set of over-expressed or under-expressed genes that are involved in the rule (number of samples in the data set that contain the gene), and their confidence (number of times that the right hand side appears in records where the left hand side gene set appears). 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 over expressed too”. Theoretically, the association rule method is able to resolve the drawbacks of existing clustering approaches. Any gene can be assigned to any number of rules as long as its expression fulfills the assigned criteria. In another words, a gene involved in many co-expression groups will appear in each of those groups. This method fits in the real-life biological processes in which one gene may be involved in many biological processes and pathways.

However, the association rule method can only capture gene co-expression, but not interactions. For example, it can not discover the rule such as “a gene is over (or under) expressed only if several genes are jointly over (under) expressed, but not if at least one of them is not over (under) expressed”. In other words, the association rule is unable to discover the interactions between different genes.

In this paper, we apply loglinear modeling, a methodology for approximating discrete multidimensional probability distributions, to discover the k-gene interactions. The remainder of the paper is structured as follows. In Section 2, we discuss existing work on association rules and loglinear modeling. In Section 3, we formally introduce how to screen k-gene interactions by using all k-way interaction loglinear models. In Section 4, we present experimental results over yeast data and give interpretation. We present the conclusion in Section 5.

2. Related Work

Both association rule 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 (Friedman et al., 2000; Murphy & Mian, 1999; Segal et al., 2003). The advantage of bayesian network is that it generates a directed graph that suggests causal influence. However, bayesian network can-

not 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.

In (Creighton & Hanash, 2003), association rules (Agrawal et al., 1993; Agrawal & Srikant, 1994) 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 (Silverstein et al., 1998), pairwise associations (DuMouchel & Pregibon, 2001) have been investigated to overcome the limitations of support-based association algorithms. For example, the authors (DuMouchel & Pregibon, 2001) selected the multi-item associations that cannot be explained by the pairwise associations in the item set by using the standard statistical theory of log-linear models. 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 discovered by other techniques (e.g., association rule (Agrawal et al., 1993), bayesian network (Heckerman, 1997), graphical gaussian model (Lauritzen, 1996)). Besides, our model can also interpret the interestingness of associations by examining loglinear parameters. We believe *k*-way gene interaction effects can significantly contribute towards the biological annotations of genes including GENMAPP (Dahlquist et al., 2002), Gene Ontology (Ashburner et al., 2000), etc.

3. Our Method

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.

Our method involves first finding large gene sets by using Apriori algorithm, building all k-way interaction model iteratively, and screening large gene sets based on the estimates from k-way interaction model. The method can be sketched as follows:

- Step 1. Transforming gene expression raw data to build a boolean matrix.
- Step 2. Apply Apriori method to find all large gene sets $\mathcal{S}^{(0)}$.
- Step 3. 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 1 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.

$$e^{(k)} = \frac{y - \hat{y}^{(k)}}{\sqrt{\hat{y}^{(k)}}} \quad (1)$$

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.

3.1. Loglinear Model Revisited

Loglinear modeling (Andersen, 1994) 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

$$\hat{l}_{i_1 i_2 \dots i_n} = \log \hat{y}_{i_1 i_2 \dots i_n} = \sum_{G \subseteq \{d_1, d_2, \dots, d_n\}} \gamma_{(i_r | d_r \in G)}^G \quad (2)$$

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 3 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 .

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

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 discover 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, $\gamma_i^{AB} = \sum_{j=0}^{J-1} \gamma_{ij}^{AB}$, where a dot “.” means that the parameter has been summed over the index. Equation 4 shows how to compute the coefficients in a 4-dimensional table.

$$\begin{aligned} \gamma &= l_{\dots} \\ \gamma_i^A &= l_{i\dots} - \gamma \\ &\dots \\ \gamma_{ij}^{AB} &= l_{ij..} - \gamma_i^A - \gamma_j^B - \gamma \\ &\dots \end{aligned} \quad (4)$$

In (Sarawagi et al., 1998) 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.

Table 1. Example matrix of gene-expression data. The rows denote samples or conditions while the columns denote genes. $G(\beta, B)$ denotes the quantitative expression of gene B in the sample β

	A	B	C
α	0.23	0.1	-0.24
β	0.6	0.1	0.5
γ	0.3	0.13	0.28
δ	0.15	0.30	-0.25
ϵ	0.8	0.08	0.30
λ	-0.2	0.5	0.25

3.2. Preprocessing Raw Data

The data in the standard association rule mining are in the form of a large boolean matrix. In the case of gene expression data, we need to discretize gene expression values to categories. For example, the values may be discretized into two categories, underexpressed or overexpressed, according to their expression levels under an experimental situation comparing with the control situation (Becquet et al., 2002)¹.

In (Creighton & Hanash, 2003), the expression values are discretized into three categories, underexpressed, normal, or overexpressed. The “normal” state of gene expression which means as being neither up nor down can effectively decrease the effect of noises. In our experiment, we apply the same discretization strategy as (Creighton & Hanash, 2003) for comparison. As each gene has three states, we map each gene as two items (one means gene overexpressed and the other means underexpressed) in the standard association rule framework. Table 1 and 2 shows one example of raw data and transformed binary matrix respectively.

It is important to note 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, e.g., underexpressed and over-expressed the contingency table built by 7 genes has 128 (2^7) cells which require more than 128 samples.

¹The control expression level of a gene can be either determined experimentally, or it can be set as the average expression level of the gene across experiments.

Table 2. Transformed boolean matrix where the thresholds are 0.2 and -0.2 respectively

	A \uparrow	A \downarrow	B \uparrow	B \downarrow	C \uparrow	C \downarrow
α	1					1
β	1				1	
γ	1				1	
δ			1			1
ϵ	1				1	
λ		1	1		1	

3.3. All k-way Interaction Loglinear Model Fitting

The Apriori method is applied here to extract all gene sets whose frequencies exceed support threshold. For each large gene set, we build one contingency table which will be analyzed by all k-way loglinear models. Table 3 shows one contingency table built from yeast data based on large item sets with four genes (e.g., YHR071W, YMR094W, YMR096W, YMR095C). 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.

$$\log \hat{y}_{ijkl}^{(1)} = \gamma + \gamma_i^A + \gamma_j^B + \gamma_k^C + \gamma_l^D \quad (5)$$

$$\log \hat{y}_{ijkl}^{(2)} = \gamma + \gamma_i^A + \gamma_j^B + \gamma_k^C + \gamma_l^D + \gamma_{ij}^{AB} + \gamma_{ik}^{AC} + \gamma_{il}^{AD} + \gamma_{jk}^{BC} + \gamma_{jl}^{BD} + \gamma_{kl}^{CD} \quad (6)$$

$$\log \hat{y}_{ijkl}^{(3)} = \gamma + \gamma_i^A + \gamma_j^B + \gamma_k^C + \gamma_l^D + \gamma_{ij}^{AB} + \gamma_{ik}^{AC} + \gamma_{il}^{AD} + \gamma_{jk}^{BC} + \gamma_{jl}^{BD} + \gamma_{kl}^{CD} + \gamma_{ijk}^{ABC} + \gamma_{ijl}^{ABD} + \gamma_{ikl}^{ACD} + \gamma_{jkl}^{BCD} \quad (7)$$

Equation 5 and 6 shows all 1-way and all 2-way interaction model respectively. Equation 5 assumes the independence model and includes all-one-factor (main) effects and grand mean. Equation 6 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 (Silverstein et al., 1998) or (Du-Mouchel & Pregibon, 2001) respectively. However, the assumed independence model or pairwise model may be insufficient to fit some gene interactions. In (Du-Mouchel & Pregibon, 2001), 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

Table 3. 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	0
	$C \uparrow$	0	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	0
	$C \uparrow$	0	0	1	0	7	2	0	1	2
$D \uparrow$	$C \downarrow$	0	0	0	0	0	0	0	0	0
	C	0	0	0	1	11	0	0	0	1
	$C \uparrow$	0	0	0	0	15	3	0	19	54

7). 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.

3.4. Interpreting Interactions by Examining Parameters

If the gene expression values are discretized to 2 categories, over-expressed and under-expressed, our previous results (Wu et al., 2003a) 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 1 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, $\gamma^{AB} = -0.044$ in Figure 1 implies $\gamma_{00}^{AB} = -0.044$, $\gamma_{01}^{AB} = 0.044$, $\gamma_{10}^{AB} = 0.044$, and $\gamma_{11}^{AB} = -0.044$. It can be interpreted that the overexpression (underexpression) of A implies the overexpression (underexpression) of B with interaction effect of 0.044. Furthermore, the comparison of γ^{AC} (0.681) and γ^{CD} (0.245) implies the interaction of AC is more significant than that of CD.

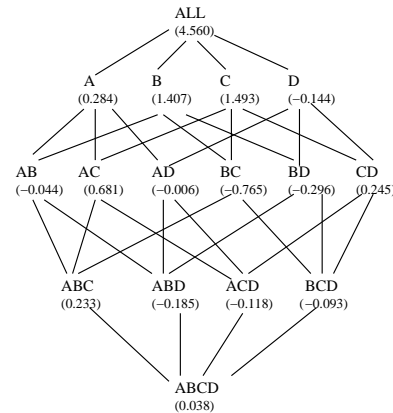


Figure 1. Lattice for the data set with four dimensions denoted by A, B, C, D respectively. The value in () denotes the value of γ -term of saturated loglinear model

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. It is important to point out that 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). Secondly, the variance for each interaction varies (in two-category case, the variances for all γ -terms are the same). The values $\gamma_{00}^{AC} = 0.681$ and $\gamma_{00}^{CD} = 0.245$ do not necessarily imply that the interaction of AC is greater than that of CD, since the variances of γ_{00}^{AC} , γ_{00}^{CD} can be different.

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 (Goodman, 1971).

3.5. Discussion of Future Work

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 graphical gaussian models, which assume a jointly normal distribution, were applied for gene expression analysis in (Kishino & Waddell, 2000; Wu et al., 2003b). The independence graph generated by graphical gaussian modeling indicates only pairwise gene interactions, and is insufficient for pathway based analysis, which require understanding higher order relationships. 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 models, 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.

The application of graphical gaussian modeling is constrained by the sample size as the correlation matrix is inevitably degenerate. We are investigating a framework consisting the following phases:

- Preprocessing: Microarray expression data is input to hierarchical clustering or association rule mining, resulting in a set of gene clusters.
- Subsets of genes (clusters or frequent itemsets) are analyzed for pairwise gene interaction using graphical gaussian models.
- The independence graph from graphical gaussian models is decomposed to obtain components. The genes included in each component are then analyzed to get higher order effects using loglinear models.
- Interactive Visualization/Analysis: The user may interactively analyze, modify and explore the output of both graphical gaussian models and loglinear models.

4. Experimental Results

In this section we show the results of experimenting with one real data set and some synthetic data sets. The experiments were conducted in a DELL PowerEdge 4400, with one 1G processor, and 1G bytes of RAM.

4.1. Data Sets

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. In (Creighton & Hanash, 2003), this 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.

4.2. Results and Interpretation

Table 4 shows size of gene sets using association rule and 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 sets ². .

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)/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 (Padilla et al., 1998).
- Our results show YER175C has interaction with SNZ1/SNO1/CTF13. YER175C encodes the trans-aconitate methyltransferase (Tmt1p) of *Saccharomyces cerevisiae*, which is localized in the cytosol and increases markedly as cells undergo the metabolic transition at the diauxic shift (Cai et al., 2001).
- CTF13, SNO1 and SNZ1, located adjacent to each other, are situated proximal to the centromere on the right arm of chromosome XIII. We project

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

Table 4. Size of gene sets obtained using association rule and k-way interaction model with different support. $\|S^{(0)}\|$ denotes the size of large item sets from Apriori method, $S^{(k)}$ ($k = 1, 2, 3$) 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 sets. 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

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 (Kinney & Lusty, 1989). 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.

5. Conclusions

In this paper we have applied a combination of association rule mining and loglinear modeling to find meaningful interactions among sets of genes in gene expression data collected by microarrays. The key to the method is to find sets of genes whose support exceeds by more than a threshold the value estimated by a k-way interaction loglinear model. We have shown that the application of the method to yeast microarray data uncovers a set of interactions that can be explained us-

ing biological arguments, and thus are meaningful. As such, we believe that this method complements the typical clustering approaches used to analyze microarray data.

References

- Agrawal, R., Imilienski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD International Conference on Management of Database* (pp. 207–216).
- Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the International Conference on Very Large Data Bases* (pp. 487–499).
- Andersen, E. (1994). *The statistical analysis of categorical data*. Springer Verlag, Berlin, Heidelberg.
- Ashburner, M., Ball, C., Blake, J., & et al. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25, 25–29.
- Becquet, C., Blachon, S., Jeudy, B., Boulicaut, J.-F., & Grandrillon, O. (2002). Strong-association-rule mining for large-scale gene-expression data analysis:

- a case study on human sage data. *Genome Biology*, 3, 0067.1–0067.16.
- Ben-Dor, A., Shamir, R., & Yakhini, Z. (1999). Clustering gene expression patterns. *Journal of Computational Biology*, 6, 281–297.
- Cai, H., Dumlao, D., Katz, J., & Clarke, S. (2001). Identification of the gene and characterization of the activity of the trans-aconitate methyltransferase from *saccharomyces cerevisiae*. *Biochemistry*, 40, 13699–13709.
- Creighton, C., & Hanash, S. (2003). Mining gene expression databases for association rules. *Bioinformatics*, 19-1, 79–86.
- Dahlquist, K., Salomonist, N., Vranizan, K., Lawlor, S., & Conklin, B. (2002). Genmapp, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet*, 31, 19–20.
- DuMouchel, W., & Pregibon, D. (2001). Empirical bayes screening for multi-item association. *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 67–76).
- Friedman, N., Linial, M., Nachman, I., & Peer, D. (2000). Using bayesian networks to analyze expression data. *Proceedings of the fourth Annual International Conference on Computational Molecular Biology*.
- Goodman, L. (1971). The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics*, 13, 33–61.
- Hartuv, E., & Shamir, R. (2000). A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76, 175–181.
- Heckerman, D. (1997). Bayesian networks for data mining. *Data Mining and Knowledge Discovery*, 1, 79–119.
- Hughes, T., Marton, M., Jones, A. R., Roberts, C., Stoughton, R., Armour, C., Bennett, H., Coffey, E., Dai, H., He, Y., Kidd, M. J., & King, A. M. (2000). Functional discovery via a compendium of expression profiles. *Cell*, 102, 109–126.
- Kinney, D., & Lusty, C. (1989). 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.
- Kishino, H., & Waddell, P. J. (2000). Correspondence analysis of genes and tissue types and finding genetic links from microarray data. *Genome Informatics*, 11, 83–95.
- Lauritzen, S. (1996). *Graphical models*. Oxford University Press.
- Murphy, K., & Mian, S. (1999). Modeling gene expression data using dynamic bayesian networks. *Technical Report, CS dept., University of California at Berkeley*.
- Padilla, P. A., Fuge, E. K., Crawford, M. E., Errett, A., & Werner-Washburne, M. (1998). The highly conserved, coregulated *sno* and *snz* gene families in *saccharomyces cerevisiae* respond to nutrient limitation. *Journal of Bacteriol*, 180, 5718–5726.
- Sarawagi, S., Agrawal, R., & Meggido, N. (1998). Discovery-driven exploration of olap data cubes. *Proceedings of the International Conference on Extending Data Base Technology* (pp. 168–182).
- Segal, E., Shapira, M., Regev, A., Peer, D., Botstein, D., Koller, D., & Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34, 166–176.
- Shamir, R., & Shamir, R. (2000). Click: A clustering algorithm for gene expression analysis. *Proceedings of the Eighth International Conference on Intelligent System for Molecular Biology (ISMB00)*.
- Silverstein, C., Brin, S., & Motwani, R. (1998). Beyond market baskets: generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 2, 39–68.
- Wu, X., Barbará, D., & Ye, Y. (2003a). Screening and interpreting multi-item associations based on log-linear modeling. *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
- Wu, X., Ye, Y., Subramanian, K., & Zhang, L. (2003b). Interactive gene interaction analysis using graphical gaussian models. *The 3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics*.
- Xu, Y., Olman, V., & Xu, D. (2002). Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics*, 18, 536–545.