

Interactive Analysis of Gene Interactions Using Graphical Gaussian Model

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

Kalpathi R. Subramanian
UNC Charlotte
9201 University City Blvd.
Charlotte, NC 28269
krs@uncc.edu

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. Association rule mining and loglinear models have been used for this purpose, but their inherent limitations as well as information loss due to discretization limit the applicability of the results. Here we propose the use of a *Graphical Gaussian Model* to discover pairwise 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 Gaussian Modeling, Gene Interaction Analysis

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). The challenge is to rapidly and efficiently extract useful information and discover knowledge from the data, such as gene functions, gene interactions, regulatory pathways, metabolic pathways, and effects of environmental factors.

Clustering algorithms (e.g., CAST [3], MST [31], HCS [10], CLICK [24]) have been quite successful in the molecular profiling of human cancers. Gene clusters from these methods can be interpreted as a network of co-regulated genes, which may encode interacting proteins that are involved in

the same biological processes such as cell cycle, metabolic pathway, signaling transduction pathway, and genetic regulatory pathway. However, clustering methods cannot identify molecular networks or analyze high level function, i.e., the gene expression changes in the context of biological pathways; this is because the clustering methods do not take into account relationships between genes within each cluster and those across different clusters. Additionally, clustering techniques assign a gene to a single cluster, while it is known that a gene, such as the p53 protein, can function in multiple physiological pathways. Therefore, there is a great need for new tools to perform gene interactions and pathway-based analyses of gene expression data, which can present the knowledge embedded in the microarray data in a manner that is intuitive and familiar to biologists.

Association rule mining [4; 2] and k-way interaction loglinear modeling [30] have been investigated for identifying gene interactions. To apply association rule or loglinear modeling we need to discretize the gene expression values into expression categories, e.g., under-expressed and over-expressed, depending on whether the expression level is significantly lower than, or higher than control¹. It is clear that by discretizing the measured expression levels we lose information. Also, as the number of genes significantly exceeds the number of samples, it may be inaccurate to apply association rule (where the number of items is assumed far less than the number of transactions) or loglinear modeling (where the size of samples is expected to be five times as large as the size of cells in contingency tables).

In this paper we study gene interactions using Graphical Gaussian Models (GGMs) which assume a family of normal distributions for underlying data constrained to satisfy the pairwise conditional independence restrictions inherent in the independence graph. It is clear that this method does not suffer from the information loss caused by discretization. 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. We also present our framework of interactive gene interaction analysis prototype system. The core components of the system is pairwise gene interactions using GGM and multiple-way gene interactions using loglinear modeling. We subject the input data of GGM to the output of other data mining techniques (e.g., clusters from hierarchical cluster-

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.

ing, frequent item sets from association rule mining), prior to analyzing gene interactions. Our system enables domain users to interactively explore gene interactions by adding or removing genes based on domain knowledge.

The remainder of the paper is structured as follows. In Section 2, we formally introduce how to analyze gene pairwise interactions using GGMs and present a prototype system for interactive gene interaction analysis. In Section 3, we present experimental results based on published yeast data and provide interpretation. The conclusion and future work are summarized in Section 4.

2. 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 . Our goal here is to identify the interactions among subsets of genes which can be discovered by other data mining techniques or specified by domain users.

In Section 2.1 we present GGMs and formalize partial correlations. Here we assume the number of genes in each subset is less than the size of samples. In Section 2.2 we present our interactive interaction analysis framework. We compare GGMs with other graphical models such as bayesian networks and loglinear modeling in Section 2.3.

2.1 Graphical Gaussian Models

Graphical gaussian model [19; 27], also known as covariance selection model, assumes multivariate normal distribution for underlying data and satisfies the pairwise conditional independence restrictions which are shown in the independence graph of a jointly normal set of random variables. 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, and provides better evidence for regulatory genetic links than standard correlation. For example, Figure 1 shows the correlation and partial correlation

graph over a subset of genes. We omit the values for pairwise correlation and pairwise partial correlation due to space limitation. Figure 1(b) shows pairwise correlations with correlation coefficient greater than 0.65, which indicates positive correlations between any pair of genes. However, partial correlations in Figure 1(a) indicates 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. The partial correlation agrees with biological interpretation.

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 [27]. 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 $pr_{x_ix_j.g}$.
- 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.
- Fitting GGMs by maximum likelihood estimation.

The core of the method is to compute the inverse of covariance matrix. We apply singular value decomposition (SVD) to compute the inverse of matrix in our prototype system. The SVD method decomposes a $m \times n$ matrix \mathbf{X} into two orthogonal matrices \mathbf{U} , \mathbf{V} and a diagonal matrix $\mathbf{\Lambda}$ where $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_n$ and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n)$, $\lambda_i > 0$ for $i = 1, \dots, r$, $\lambda_i = 0$ for $i = r + 1, \dots, n$. The SVD has the optimal truncation property: if we discard all but the r largest singular values and the corresponding singular vectors, the product of $\mathbf{U}'\mathbf{S}'\mathbf{V}'$ is the best rank- r approximation of \mathbf{X} in the least-squares sense. In general, SVD needs $O(mn)$ space and $O(nm^2)$ (or $O(mn^2)$ depending on which one is smaller) computation.

2.2 Interactive Analysis of Gene Interactions

Our goal is to explore inter-relationships among genes. To make this process intuitive and efficient, we propose to integrate interactive techniques and information visualization to interaction modeling. Figure 2 shows the framework of our proposed prototype system for interactive gene interaction analysis, consisting of the following three major phases:

- **Preprocessing:** Microarray expression data is input to hierarchical clustering or association rule mining, resulting in a set of gene clusters.
- **Data Modeling:**
 - Subsets of genes (clusters or frequent itemsets) are analyzed for pairwise gene interaction using graphical gaussian models.

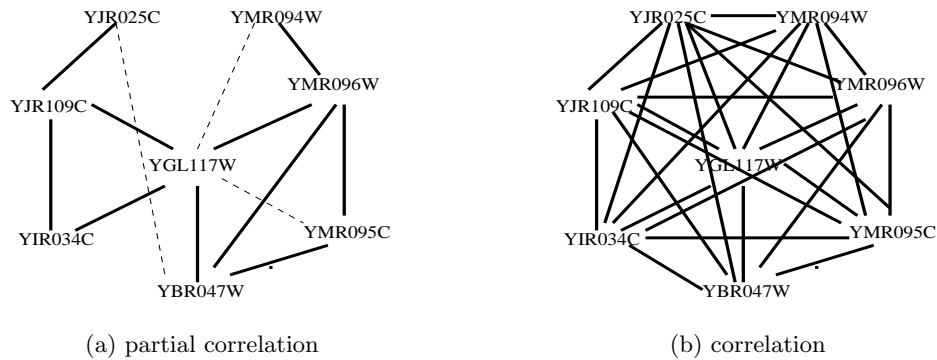


Figure 1: Gene interactions using partial correlation vs. correlation, the threshold for partial correlation is 0.2 while the threshold for correlation is 0.65. Note dashed lines indicate a negative partial correlation and solid lines indicate a positive partial correlation.

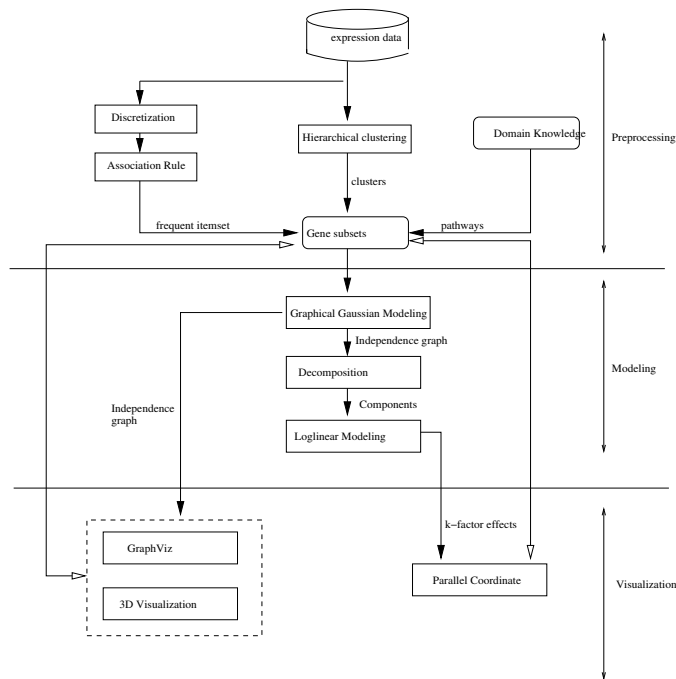


Figure 2: The framework of prototype system of gene interaction analysis

- 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.

The microarray data is preprocessed via hierarchical clustering or association rule mining prior to analyzing gene pairwise interactions using GGMs, as (1) the large data size makes it infeasible to apply the GGMs, and (2) the correlation matrix is generally degenerate, as the matrix rank is bounded by the sample size. The number of genes contained in the resulting clusters or frequent itemsets is usually less than the size of samples, thus avoiding the matrix rank problem. The authors, in [17], propose 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. The independence graph generated by graphical gaussian modeling can give domain users a basic understanding of interactions among relatively large gene subsets. However, The independence graph indicates only pair-wise gene interactions, and is insufficient for pathway based analysis, which require understanding higher order relationships.

To extract multi-way interactions of genes, we need to apply loglinear modeling which assumes multinomial distributions (For details see [30; 29]). However 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., under-expressed and over-expressed, depending on whether the expression level is significantly lower than, or higher than control², the contingency table built by 7 genes has 128 (2^7) cells which require more than 128 samples. Hence we propose to decompose independence graph into components and apply loglinear modeling on each component. It is worth pointing out 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 [1], bayesian network [11], graphical gaussian model [19]).

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 propose the use of *interactive techniques*, whereby a user can interactively analyze gene interactions by adding or removing any number of genes to/from one subset. To make this interactive exploration intuitive and efficient, we applied information visualization techniques, whereby visual representations present the interface to interactive exploration. In this work, we use automatic graph drawing algorithms [8] to display and edit gene subsets and their 2-way relationships. We are also working on interactive visual representations for cluster hierarchies [25] as well as association rule mine sets [28], so as

²The 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.

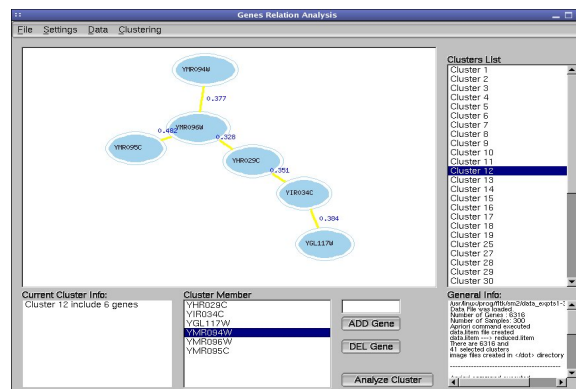


Figure 3: The snapshot of prototype system of gene interaction analysis

to rapidly focus, view and interactively edit gene subsets of interest. A log of a user's analysis session can be easily kept track of for review, or continuation from a previous session. Figure 3 shows a preliminary snapshot of our prototype system of gene interaction analysis.

2.3 Discussion

The graphical gaussian model method is statistically sound and computationally tractable for analyzing microarray data and inferring biological interactions from them. However, it can only detect dependencies that are close to linear. In particular, it is not likely to discover combinatorial effects (e.g., a gene is over expressed only if several genes are jointly over expressed, but not if at least one of them is not over-expressed). On the other hand, loglinear modeling, which assumes multinomial distribution, can reveal combinatorial effects. However, loglinear modeling can only be applied to a small set of genes due to small size of samples in microarray data. Furthermore, loglinear modeling loses information due to discretization [30]. We are currently trying to combine graphical gaussian and loglinear modeling for gene interaction analysis in our prototype system.

Both graphical gaussian 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 analysis [7; 21]. The bayesian network is a directed graph-based model of joint multivariate probability distributions that captures properties of conditional independence between variables. The problem of applying bayesian network to the analysis of microarray data is that learning the bayesian network structure is a NP-hard problem as the number of DAGs is superexponential in the number of genes and exhaustive search is intractable.

Several public and commercial resources exist for pathway based analysis, including the Alliance for Cellular Signaling, BioCarta, EcoCyc [15], MetaCyc [14], KEGG[13] and PathDB[18]. These databases contain large amounts of curated information; EcoCyc and KEGG allow viewing simple gene expression data over pre-existing pathways, and GenMAPP [5] extends the capabilities of these pathways. However, all these tools only indicate pathways currently recog-

Table 1: Size of gene sets obtained using frequent itemset and maximal frequent itemset mining with different support

support(%)	frequent itemset	maximal frequent itemset
12	130603	1635
13	22123	795
14	2735	298
15	1134	164
16	314	69
17	79	23
18	39	16
19	17	10
20	8	4
21	2	2

nized in textbooks and literature; it is beyond their capability to identify/predict new gene interactions and pathways from DNA microarray data.

3. EXPERIMENTAL RESULTS

In this section we show the results on yeast data [12] which contains expression profiles for 6316 transcripts corresponding to 300 diverse mutations and chemical treatments in yeast. We use automatic graph drawing tools [8; 6] to represent gene networks. Our implementation is in C++ on Unix workstations using FLTK [26] for the user interface.

We apply frequent item set and maximal frequent item set mining methods³ to get gene subsets. In [4], 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. Table 1 shows the size of gene sets obtained using frequent item set and maximal frequent item set mining method with different support. We can see the size of frequent item set and maximal frequent item set under low support values is large. How to further prune them while keeping relevant or interesting gene sets remains an open problem.

Figure 4 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 this gene has no correlation with the remaining genes,

³A frequent itemset is called *maximal* if it is not a subset of any other frequent itemsets. See [9] for an efficient algorithm.

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 4 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-membered 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 [22; 20]. 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 [23].

Furthermore, our results reveal some previously unknown interactions that have solid biological explanations. For example, CTF13 (centromere transmission fidelity, CTF) encoded by YM094W is an essential protein in the Cbf3 kinetochore protein complex, which binds to the centromeres during mitosis. CTF13 and SNZ1, located adjacent to each other, are situated proximal to the centromere on the right arm of chromosome XIII. We project that the correlation of the expression of these two genes might be caused by the conformational changes of chromosomal structure during transcription activation even though the possibilities that they are involved in the same biological process and/or they can directly interact with each other are not excluded. YJR109C and YIL116W encode Cpa2 and His5 which are involved in arginine and histidine biosynthesis, respectively. Arginine restriction led to increased expression of HIS3, CPA1, and CPA2 in *Saccharomyces cerevisiae*, which indicates that the regulation of arginine biosynthesis pathway is related to that of histidine biosynthesis pathway [16].

Our data also indicates that the expression of HIS5 (involved in histidine biosynthesis pathway) correlates with that of CPA2 and LYS1, which are involved in arginine and lysine biosynthesis pathway, respectively. Yhm1, encoded by YDL198C, is a transporter which resides on mitochondrial inner membrane. The correlation of YHM1 and CPA2/LYS1 indicate that Yhm1 might be a mitochondrial carrier involved in arginine and lysine biosynthesis. In addition, since the biological function of YKL218C and YOL118C are unknown, we speculate that YKL218 might be involved in NAD biosynthesis pathway since its expression correlates with that of YJR025C, whereas YOL118C might be related with vitamin B2 biosynthesis since its expression correlates with that of YBR256C (RIB5).

4. CONCLUSIONS

In this paper we have applied graphical gaussian models to find meaningful pairwise interactions among sets of genes in gene expression data collected by microarrays. Graphical gaussian modeling has the advantage of being able to model conditional distributions of continuous variables. We have

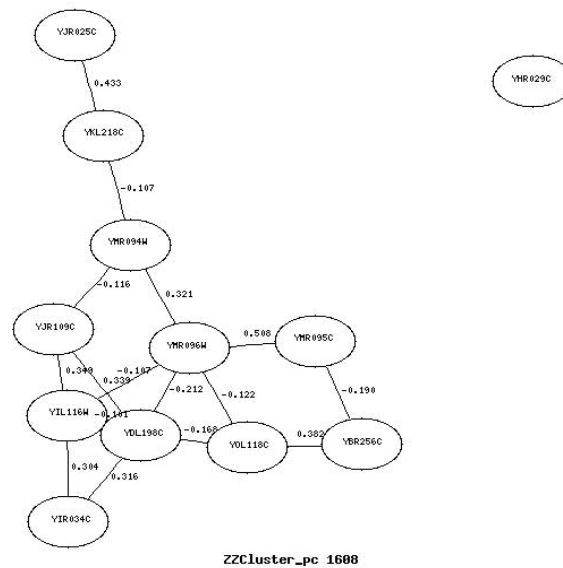


Figure 4: Pairwise gene interactions using GGMs for a selected maximal frequent gene set

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.

We also present our framework of interactive gene interaction analysis system. Our combined interaction analysis models will reveal complex gene interactions and the interactive visualization system allows efficient gene interaction analysis and pathway exploration. We expect the proposed work will complement the functionalities of currently available resources for pathway analysis, by providing a new tool for the analysis of gene interaction and genetic networks.

5. ADDITIONAL AUTHORS

Additional authors: Liying Zhang (Memorial Sloan Kettering Cancer Center, email: zhang12@mskcc.org).

6. 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] C. Becquet, S. Blachon, B. Jeudy, J.-F. Boulicaut, and O. Grandrillon. Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human sage data. *Genome Biology*, 3(12):0067.1–0067.16, 2002.
- [3] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3/4):281–297, 1999.
- [4] C. Creighton and S. Hanash. Mining gene expression databases for association rules. *Bioinformatics*, 19:1:79–86, 2003.
- [5] 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.
- [6] J. Ellson and S. North. Graph visualization project (graphviz). <http://www.graphviz.org>.
- [7] 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.
- [8] E. Gansner and S. North. An open graph visualization system and its applications to software engineering. *Software - Practice and Experience*, 30(11):1203–1233, 2000.
- [9] K. Gouda and M. J. Zaki. Efficiently mining maximal frequent itemsets. In *1st IEEE International Conference on Data Mining, San Jose, California*, Nov 2001.
- [10] E. Hartuv and R. Shamir. A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76(4-6):175–181, 2000.
- [11] D. Heckerman. Bayesian networks for data mining. *Data Mining and Knowledge Discovery*, 1:79–119, 1997.
- [12] 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.
- [13] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya. The kegg databases at genomenet. *Nucleic Acids Res*, 30:42–46, 2002.
- [14] P. Karp, M. Riley, S. Paley, and A. Pellegrini-Toole. The metacyc database. *Nucleic Acids Res*, 30:59–61, 2002.

- [15] P. Karp, M. Riley, M. Saier, and et al. The ecoecyc database. *Nucleic Acids Res*, 30:56–58, 2002.
- [16] 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 Biol.*, 9(11):4882–8, 1989.
- [17] 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.
- [18] R. Kuffner, M. Gonzales, P. Steadman, and et al. Pathdb. <http://www.ncgr.org/pathdb>, National Center for Genome Resources.
- [19] S. Lauritzen. *Graphical Models*. Oxford University Press, 1996.
- [20] G. Mittenhuber. Phylogenetic analyses and comparative genomics of vitamin b6 (pyridoxine) and pyridoxal phosphate biosynthesis pathways. *Journal of Mol. Microbiol Biotechnol*, 3(1):1–20, 2001.
- [21] K. Murphy and S. Mian. Modeling gene expression data using dynamic bayesian networks. *Technical Report, CS dept., University of California at Berkeley*, 1999.
- [22] P. A. Padilla, E. K. Fuge, M. E. Crawford, A. Erett, 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.
- [23] S. Rodriguez-Navarro, B. Llorente, M. Rodriguez-Manzanegue, A. Ramne, G. Uber, D. Marchesan, B. Dujon, E. Herrero, P. Sunnerhagen, and J. Perez-Ortin. Functional analysis of yeast gene families involved in metabolism of vitamins b1 and b6. *Yeast*, 19(14):1261–1276, 2002.
- [24] 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.
- [25] B. Shneiderman. Tree visualization with tree-maps:2-d space filling approach. *tog*, 11(1):92–99, 1992.
- [26] B. Spitzak and et. al. The fast light toolkit(ftk). <http://www.ftk.org>.
- [27] J. Whittaker. *Graphical Models in Applied Mathematical Multivariate Statistics*. Wiley, 1990.
- [28] P. Wong, P. Whitney, and J. Thomas. Visualizing association rules for text mining. In *IEEE Symposium on Information Visualization 1999 (INFOVIS'99), San Francisco, California*, pages 24–29, October 1999.
- [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*. 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*. Washington, DC, August 2003.
- [31] 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.