

GenExplore: Interactive Exploration of Gene Interactions from Microarray Data

Yong Ye, Xintao Wu, Kalpathi Subramanian
University of North Carolina at Charlotte
Dept. of Computer Science
9201 University City Blvd., Charlotte, NC 28223
{yye,xwu,krs}@uncc.edu

Liyang Zhang
Memorial Sloan Kettering Cancer Center
Dept. of Pathology
New York, NY 10021
zhang12@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 project, we propose to combine graphical model based interaction analysis with other data mining techniques (e.g. association rule, hierarchical clustering) for this purpose. For interaction analysis, we propose the use of Graphical Gaussian Model to discover pairwise gene interactions and 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.

1. Motivation

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, MST, HCS, CLICK, etc.) have been quite successful in the molecular profiling of human cancers, however they are insufficient to identify molecular networks (i.e., the structure of the transcriptional regulation process). In clustering, each gene is assigned to only one cluster. However, a gene can usually be characterized in more than one way (e.g., the p53 protein belongs to more than one physiological pathway). Furthermore, it is impossible to determine the interactions that exist between different genes within each cluster, especially in situations where a gene participates in multiple biological pathways.

We have been building a prototype system which allows user to explore and analyze gene interactions effectively and efficiently. The core of the system is gene interaction analysis using Graphical Gaussian Modeling (GGM) and log-linear modeling. We subject the input data of GGM and loglinear model to the output of other data mining techniques (e.g., clusters from hierarchical clustering, frequent item sets from association rule mining), prior to analyzing gene interactions. Our system also enables domain users to interactively explore gene interactions by adding or removing genes based on domain knowledge.

2. System Overview

Our goal is to explore inter-relationships between a subset of genes. To make this process intuitive and efficient, we propose to combine interactive techniques and information visualization with data modeling. Figure 1 shows the framework of our proposed prototype system for interactive gene interaction analysis. Specifically, it involves the following steps:

- We subject the input data to hierarchical clustering or association rule mining, prior to analyzing gene interactions.
- Subsets of genes (clusters or frequent itemsets) are then analyzed for pairwise gene interaction using GGMs.

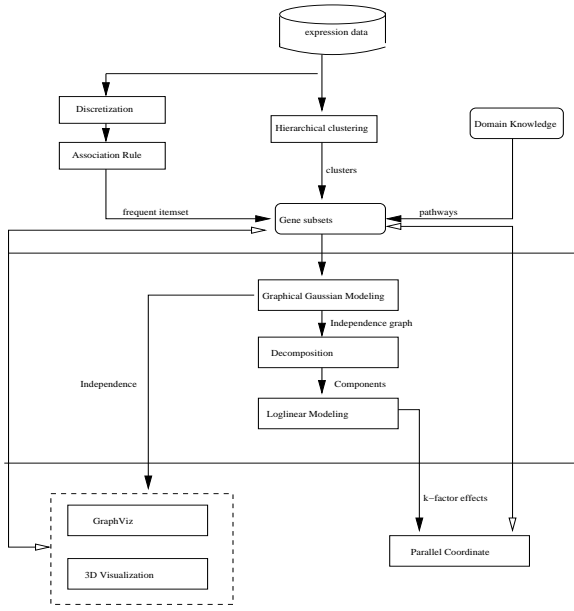


Figure 1. The framework of prototype system of gene interaction analysis

- The independence graph from graphical gaussian models is then decomposed to get components. The genes included in each component are then analyzed to get high-factor effects by using loglinear models.
- 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. In our system, we use the output of other data mining techniques. 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 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 overexpressed). To discover combinatorial effects, we need to apply loglinear modeling which assumes

multinomial distribution and requires a discretization of the data [4].

It is very likely that the number of cells in contingency table (which is determined by the number of genes and the number of categories for each gene) still significantly exceeds the number of samples, it may be inaccurate to apply for loglinear modeling directly. In our system, we decompose independence graph into components and each component is 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 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 [1] to display and edit gene subsets and their 2-way relationships and use parallel coordinate technique 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. A log of a user’s analysis session can be easily kept track of for review, or continuation from a previous session.

3. Interaction Analysis Techniques

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 . There are two models which can be used for interaction analysis: graphical gaussian models and loglinear models.

3.1. Graphical Gaussian Modeling

In this project we investigate 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 [5]. 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.

Graphical gaussian model, 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. Our result shows the partial correlation agrees with biological interpretation [1]. 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 ($S = \mathcal{V}^{-1}$). It is known that conditional independence constraints are equivalent to specifying zeros in the inverse variance [2]. 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 $S = \mathcal{V}^{-1}$.
- Scale 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.
- 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.

3.2. Loglinear Modeling

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.

For 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:

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

We will refer to the γ terms as 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-bys.

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 . 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, B . The singly-subscripted terms are analogous to main effects, and the doubly-subscripted terms are analogous to two-factor interactions. The value of each γ -term may imply the significance of interaction. By comparing different γ -terms, we can discover the patterns of combinatorial effects among a subset of genes [3, 4].

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

Equation 4 shows the linear constraints among coefficients, where a dot “.” means that the parameter has been summed over the index (For example, $\gamma_i^{AB} = \sum_{j=0}^{J-1} \gamma_{ij}^{AB}$). In short, the constraints specify that the loglinear parameters sum to 0 over all indices.

$$\gamma_i^A = \gamma_i^B = \gamma_i^C = \gamma_i^D = 0$$

$$\begin{aligned} \gamma_{i.}^{AB} = \gamma_{.j}^{AB} = \gamma_{i.}^{AC} = \gamma_{.k}^{AC} = \dots = \gamma_{.l}^{CD} = 0 \\ \dots \\ \gamma_{ijk.}^{ABCD} = \gamma_{ij.l}^{ABCD} = \gamma_{i.kl}^{ABCD} = \gamma_{.jkl}^{ABCD} = 0 \end{aligned} \quad (4)$$

Equation 5 shows how to compute the coefficients in a 4-dimensional table.

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

To apply 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.

4. Demonstration

The program has features that allow its users to choose association rule or hierarchical clustering 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. In our demonstration, we will also show the combinatorial effects from loglinear modeling using parallel coordinate techniques. Figure 2 shows a snapshot of the program. More information on the project can be found via <http://www.cs.uncc.edu/~xwu/bio/GenExplore.html>

References

- [1] 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.
- [2] J. Whittaker. *Graphical Models in Applied Mathematical Multivariate Statistics*. Wiley, 1990.
- [3] 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.

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.

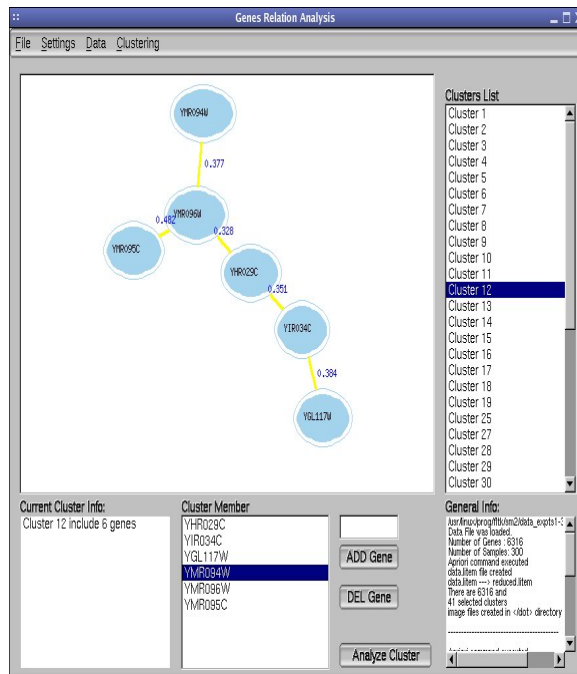


Figure 2. The snapshot of prototype system of gene interaction analysis

- [4] 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 *Proceedings of ICML Workshop on Bioinformatics in Machine Learning*. Washington, DC, August 2003.
- [5] X. Wu, Y. Ye, K. Subramanian, and L. Zhang. Interactive gene interaction analysis using graphical gaussian models. *Technical Report, UNC Charlotte*, 2003.