# GenExplore: Interactive Exploration of Gene Interactions from Microarray Data

Yong Ye, Xintao Wu, Kalpathi R. Subramanian
University of North Carolina at Charlotte
{yye,xwu,krs }@uncc.edu

Liying Zhang
Memorial Sloan Kettering Cancer Center
zhangl2@mkscc.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.*

## 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.

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 loglinear 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. First 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.

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 apply loglinear modeling which assumes multinomial distribution and requires a discretization of the data. During this process, the users may explore the output of both GGMs and loglinear models interactively given the inaccuracies and limitations of modeling methods.

## 3. 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. More information on the project can be found via http://www.cs.uncc.edu/ xwu/bio/GenExplore.html

Proceedings of the 20th International Conference on Data Engineering (ICDE'04)
1063-6382/04 $ 20.00 © 2004 IEEE