

Disease Gene Explorer: Display Disease Gene Dependency by Combining Bayesian Networks with Clustering

Qian Diao, We Hu, Hao Zhong, Juntao Li, Feng Xue, Tao Wang, Yimin Zhang
Intel China Research Center, Beijing 100020
{qian.diao, wei.hu, hao.zhong, tao.wang, yimin.zhang}@intel.com

Abstract

Constructing gene networks is one of the hot topics in the analysis of the microarray gene expression data. When combined with the output of disease gene finding, the generated gene networks will give a recommendation mechanism and an intuitive form for biologists to identify the underlying relationship among those biomarkers of the disease. In this paper, we present a display system, Disease Gene Explorer, which can graphically display the dependency among genes, especially those biomarkers of a disease. It combines Bayesian networks (BN) learning with clustering and disease gene selection. We test the system on Colon cancer data set and obtain some interesting results: most high-score biomarkers of the disease are partitioned into one group; the dependency among these disease genes are displayed as a directed acyclic graph (DAG).

1. Introduction

The typical cluster analysis [1] [3] can not give complicated structural relationship for genome-wide expression data from DNA microarray hybridization, for example, hierarchical clustering is only able to describe tree-like structure. Therefore, recently Bayesian network (BN) becomes an attractive method for constructing genetic networks from a graph-theoretic approach [4]. Methods for learning Bayesian networks can discover

dependency structure between genes. Unfortunately, in complex domains, such as gene expression, the amount of data is rarely enough to robustly learn a BN model of the underlying distribution. In such situations, statistical noise is likely to lead to spurious dependencies, resulting in models that significantly overfit the data [5].

In this paper we make an effort to solve the problem by clustering expression data before making BN learning. Module network in [5] is a tight combination between BN learning and a clustering. Our approach is a loose one, which makes clustering and BN structure learning separately instead of iteratively. As a result, we implement a system, Disease Gene Explorer, to graphically display the relationship among genes, especially those biomarkers of a disease.

2. System

The System consists of three parts, which are clustering, BN learning and disease gene finding. The architecture is illustrated as Figure 1. In order to reduce possible errors from clustering phase, we designed a novel clustering algorithm based on K-means and t-test, in which after finding the nearest cluster for a gene, we add a hypothesis testing to verify the statistical significance of correlation between them. The t-test K-means performs overall better than standard K-means and is less or even not dependent on the initial partition in both temporal gene expression

and time-course gene expression data [2]. There are two BN learning usages in the system, one is for genes inside a cluster, and the other is across the different clusters. For the former, the training set is $\{g_1, \dots, g_{|c_k|}\}$, where $|c_k|$ is the number of genes in the cluster k ; for the later, we applied module network algorithm of [5] on the discretized data. The BN learning part is implemented by Intel's Probabilistic Networks Library (PNL) (<http://www.intel.com/research/mrl/pnl/>).

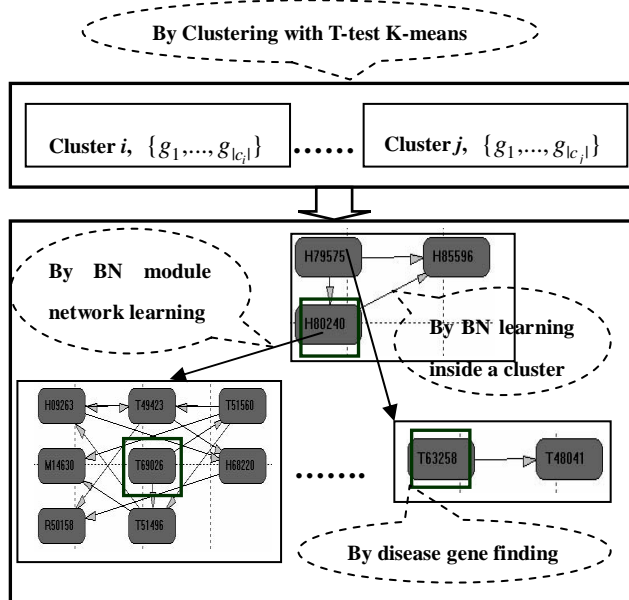


Figure 1 Architecture of disease gene explorer

3. Discussion

We test the system on Colon data. The data set consists of expression profiles for 2000 genes using an Affymetrix Oligonucleotide array in 22 normal and 40 cancer colon tissues, which were originally downloaded from <http://www.molbio.princeton.edu/colondata> [6].

A part of result is shown in Figure 2. An interesting thing is that high-score biomarkers for colon cancer referred in [6] are mostly clustered in one group by t-test K-means, and Human mRNA for snRNP E protein (X12466) and Human HF.12 gene mRNA (07290) has causal relations to the group. BN learning in our system is quite benefit for this point. The BN learning results

inside the cluster give a directed acyclic graph (DAG) to show the details of the dependency of the genes inside the cluster. The BN module network learning results give the details of on which genes the cluster has dependency.

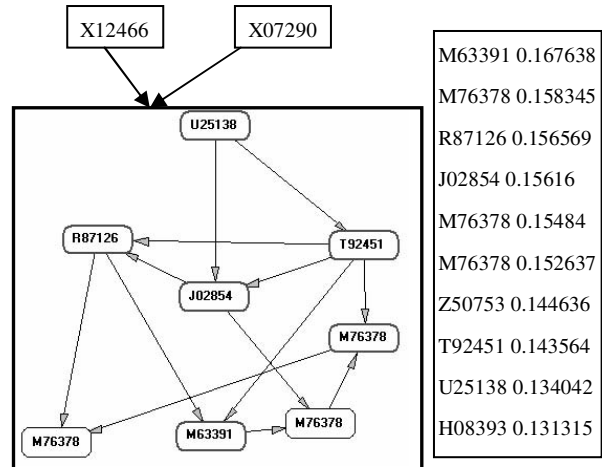


Figure 2 A part of result of disease gene explorer, the right column shows the best 10 genes selected by the t-test measure

References

- [1] Alon U., Barkai N., Notterman D. A., Gish K., Ybarra S., Mack D., and Levine A. J. (1999) Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays, *Proc. Natl. Acad. Sci. U.S.A.* 96(12), 6745-6750.
- [2] Diao Q., Zhang C., Zhang Y., Hu W., Wang T., Zhang X., (2004), A T-test K-means Approach to Gene Clustering, Technical Report, Intel.
- [3] Eisen M., Spellman P., Brown P., Botstein D. (1998) Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. U.S.A.* 95, 14863-14868.
- [4] Friedman N., Linial M., Nachman I., D. Pe'er, (2000), Using Bayesian Networks to Analyze Expression Data, *Journal of Computational Biology*, 7:601-620, 2000.
- [5] Segal, E., Pe'er D., Regev A., Koller D., Friedman N., (2003) Learning Module Networks, *Proc. of UAI*, 525-534.
- [6] Xiong Momiao, Fang Xiangzhong, and Zhao Jinying (2001) Biomarker Identification by Feature Wrappers, *Genome Research*, 11(11), 1878-87