Supervised Model for Identifying Differentially Expressed Genes in DNA Microarray Gene Expression Dataset Using Biological Pathway Information

Tae Su Chung[1,2], Keewon Kim[1] and Ju Han Kim[1,2]*

[1]Seoul National University Biomedical Informatics (SNUBI), [2]Human Genome Research Institute, Seoul National University College of Medicine, Seoul 110-799, Korea

Abstract

Microarray technology makes it possible to measure the expressions of tens of thousands of genes simultaneously under various experimental conditions. Identifying differentially expressed genes in each single experimental condition is one of the most common first steps in microarray gene expression data analysis. Reasonable choices of thresholds for determining differentially expressed genes are used for the next-step-analysis with suitable statistical significances. We present a supervised model for identifying DEGs using pathway information based on the global connectivity structure. Pathway information can be regarded as a collection of biological knowledge, thus we are trying to determine the optimal threshold so that the consequential connectivity structure can be the most compatible with the existing pathway information. The significant feature of our model is that it uses established knowledge as a reference to determine the direction of analyzing microarray dataset. In the most of previous work, only intrinsic information in the miroarray is used for the identifying DEGs. We hope that our proposed method could contribute to construct biologically meaningful structure from microarray datasets.

Keywords: biological pathway, differentially expressed gene, graph theory, DNA microarray, gene expression

Introduction

Microarray expression datasets are incessantly accumulated with the aid of recent technological advances. It is widely believed that biological meaningful interpretation can be extracted from these large-scale data, using suitable and well-organized methods of analysis. Identifying differentially expressed genes in each single experimental condition is common first step for the DNA microarray data analysis.

Newton *et al.* (Newton *et al.*, 2001) suggested the probability model to estimate the significant changes of each probe, considering not only its expression ratio of two channels but also its variance and mean intensity of two channels. To give significant genes that discriminate given groups of experimental conditions, Iyer *et al* (Iyer *et al.*, 1999) and DeRisi *et al.* (DeRisi *et al.*, 1997) discussed methods choosing genes that have big fold-changes, with suitable threshold, over their base line expressions. Also several probability approaches have been proposed to detect differentially expressed genes. (Tusher *et al.*, 2001; Hunter *et al.*, 2001; Dudoit *et al.*, 2002; Efron and Tibshirani, 2002; Zhao and Pan, 2002) But all these methods, in practical use, need to select the optimal threshold determining differentially expressed genes.

On the other hands, recently there are many attempt to lead a new insight by combining multiple types of data. (Yamanishi *et al,*. 2004; Kharchenko *et al.*, 2004) Yamanishi *et al.* uses kernel method to predict new gene-to-gene interaction within metabolic pathway and bases it on known pathway knowledge by adopting supervised information acquired from microarray expression data. Kharchenko *et al.* compares established metabolic network with expression profiles to find genes that can complete a metabolic pathway with some participants missed.

In this paper, we present a supervised model for identifying differentially expressed genes in each condition by taking optimal threshold using pathway information based on the global connectivity structure. Pathway information can be regarded as a collection of biological knowledge, thus we are trying to determine the optimal threshold so that the consequential connectivity structure can be the most compatible with the existing pathway information. Most of previous models that shared with our goal used only intrinsic information in the miroarray expression data. The significant feature of our model is that it uses established knowledge as a reference to determine the direction of analyzing

*Corresponding author: E-mail juhan@snu.ac.kr, Tel +82-2-740-8320, Fax +82-2-747-4830
Accepted 30 January 2005

microarray dataset. We hope that our proposed method could contribute to construct biologically meaningful structure from microarray data sets.

Methods

Graph structure as a common template structure

In order to compare structures of two heterogeneous types of data, microarray expression data and biological pathway data, we introduce the common template structure, *a graph*, which involves the interactive information of genes from two sources of data.

We use the Rosetta compendium dataset (Hugh *et al.*, 2000), which is hitherto the most systematic approach to profile yeast genes, as a source of microarray expression data to analysis. The dataset is consisted of 300 microarray experiment results, which contain 287 diverse gene mutations and 13 chemical treatments. They all cover 6,153 genes in each microarray data. The log-expression ratio values are used as entries of expression matrix, and these values are normalized so that mean and standard deviation of each column are 0 and 1, respectively.

In each experiment, we are trying to identify differentially expressed genes (or DEGs), which are usually determined by genes whose expression levels (or its suitable statistics) exceed some threshold. We here note that once DEGs are determined, the graph structure on the whole genes is naturally introduced by linking co-differentially expressed genes. Here by co-differently expressed genes we mean that they are DEGs under the same experimental condition. In this paper, we find optimal thresholds in the sense that the resulting secondary graph structure is most similar to the graph constructed from pathway knowledge.

KEGG (Kyoto Encyclopedia of genes and genomes) database (Kanehisa, 1996) is taken as the source of pathway knowledge. Note that it provides 88 biological pathways including 84 metabolic pathways and 4 regulatory pathways. Among these 88 pathways, we select 43 pathways that include 12 or more genes to avoid the perturbation caused by scarcity of basis knowledge. KEGG database presents a pathway with participating genes and relations between them.

In constructing 43 pathway graphs from the information, we make a node for each gene and link a pair of nodes when they are assigned one of the relations listed above. Merging these 43 pathways graphs, we build the single pathway graph of 570 genes as nodes, which will be used as a template to determine DEGs.

The global compatibility between two graphs

Here we introduce the notion of compatibility between two graphs, in the general context of graph theory. The geodesic distance $d(g,h:G)$ between two nodes g and h is defined by the length of shortest path between two nodes in the graph G. This distance represents the global structure of graph. (Chatrand *et al.*, 1998) If two graphs G1 and G2 are constructed on the same set of nodes, then the geodesic distance of two graphs can be easily extended by the average of differences of all geodesic distances of all pairs of nodes in each graph, i.e.

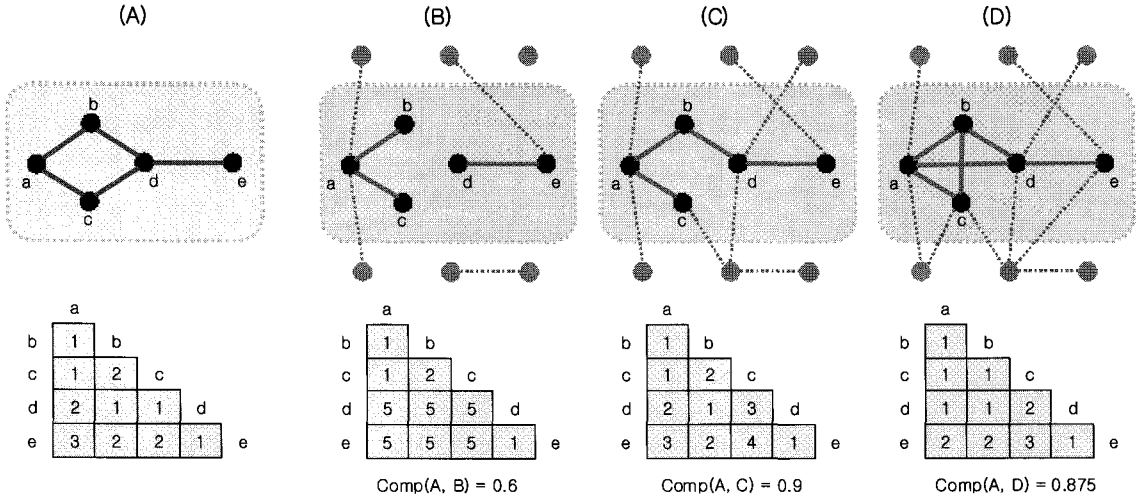

Fig. 1. Compatibilities between pathway graph and various graphs from microarray data: (A) is the pathway graph and (B)-(D) are graphs from microarray data with various thresholds. Triangular tables represent geodesic distances in graph (A) and subgraph of (B)-(D), respectively.

$$dist(G_1, G_2) = \frac{\sum |d(g,h;G_1) - d(g,h;G_2)|}{n(n-1)/2},$$

where the summation is taken over all pairs (g, h) of nodes and n is the number of common set of nodes. We here note that it is symmetric and satisfies triangle inequality (Chatrand *et al.*, 1998).

Let G_P be the pathway graph obtained from KEGG pathway database. And let G_M be a graph constructed from the Rosetta compendium dataset by linking co-differentially expressed genes. The threshold θ is used to determine DEGs in each condition, i.e., a gene pair (g,h) is linked in the graph G_M if the absolute value of normalized log-ratios of expressions in g and h are both greater than θ. Then the compatibility $Comp(G_P, G_M)$ is obtained by

$$Comp(G_P, G_M) = 1 - dist(G_M|G_P)/(n-1)$$

Here n is the number of genes in G_P, and $G_M|G_P$ is the relative subgraph of G_M to G_P. Since the pathway graph G_P contains only subset of genes that are described in graph G_M, it is natural to compare G_P and subgraph of G_M. It is clear that the compatibility lies between 0 and 1 and it becomes 1 only when the graph G_M includes exactly same structure of G_P.

The effect of the optimization on compatibility is illustrated in Fig. 1. In the figure, (B), (C) and (D) are possible graphs from microarray dataset by taking different threshold determining DEGs. And (A) is the pathway graph which will be used as a reference to select one graph from (B), (C) and (D). By calculating compatibilities, which represents the global similarity of graph structures, we can take the threshold used in constructing graph (B) as the optimal one.

Results

Correlation of micrarray data and pathway knowledge

We try to determine DEGs in each experimental condition of microarray expression data based on the template structure of pathway knowledge. The underlying assumption of our approach is that pathway knowledge can be conjectured from microarray data, or that microarray data reflect the biological pathway knowledge. We investigate the validity of this assumption by investigating the correlation between co-degree in microarray expression data and relations within pathways.

The co-degree of gene pair (g,h) with parameter θ is defined by the number of conditions satisfying that both g and h become DEGs simultaneously under the condition with respect to the threshold θ. Based on the pathway information, however, we can categorize gene pairs into three classes: The first class comprises gene pairs in which two genes belong to the pathway and are connected by a specific biological relation. The second class comprises gene pairs whose genes belong to the pathway but are not related by any biological relation. The third one is comprises gene pairs whose genes are outside the pathway. For these three classes, we calculate the average of co-degrees as a correlation of microarray data and pathway knowledge.

We naturally expected that the average co-degrees on gene pairs in the first class is bigger than that of the second class, and that of the third class the smallest value. Fig. 2 shows the result which agrees with our expectation and so we validate that there is a positive correlation between the relational information of pathway information and microarray data.

Differentially expressed genes in the Rosetta compendium dataset

Applying our model to rosetta compendium dataset, we find the optimal threshold determining DEGs in whole 300 conditions. To do this, we first let G_P be the pathway graph obtained from KEGG pathway database and G_M be a graph constructed from the rosetta compendium dataset by linking co-differentially expressed genes, where differentially expressed genes in each condition are determined by a predefined threshold. The compatibility plot via various thresholds is provided in Fig. 3. The compatibilities in the figure are calculated with two graphs G_P and G_M . The shape of the curve is nearly uni-peaked, so there is no doubt about selecting threshold in the peak as the optimal one. Actually, we just adapt simple greed algorithm to find optimal threshold 1.4926 under the restriction that the threshold is independent from conditions for the simplicity of algorithm.

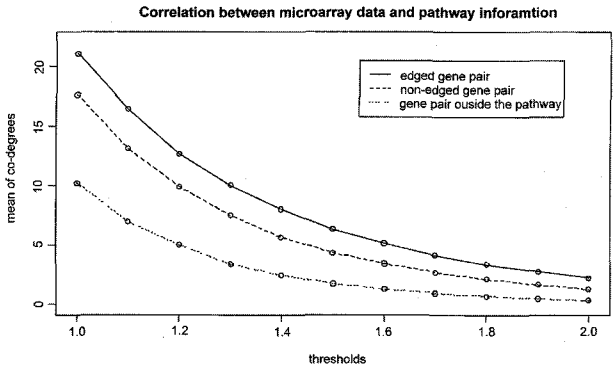

Fig. 2. Correlation between microarray data and pathway information

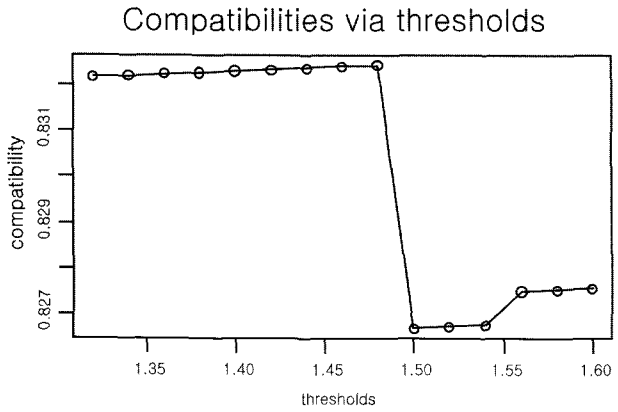

Fig. 3. Compatibilities via threshold

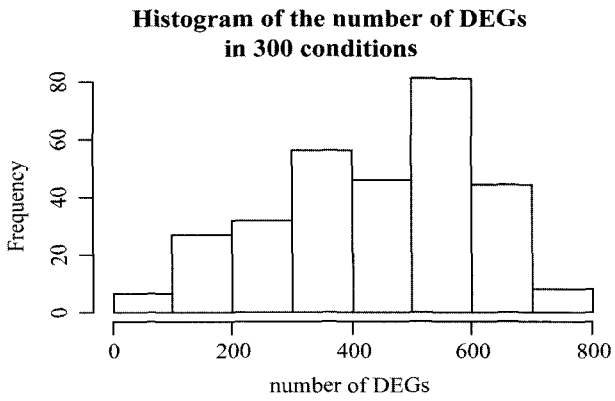

Fig. 4. Histogram of the number of DEGs in 300 conditions of the rosetta compendium datase

Fig. 4 shows the distribution of the number of DEGs in 300 conditions applying the optimal threshold. Among whole 6,153 yeast genes, we see that our model select about 10% of genes as DEGs in each condition. This result agrees with the common criterion which can be found in many biological literatures.

Discussion

In the present paper, we suggest a novel model to identify differentially expressed genes in each conditions of a microarray dataset. The procedure is one of the first steps of analyzing expression profiles. The significant feature of our model is that it uses established knowledge as a reference to determine the direction of analyzing microarray dataset and it does not consider the individual structure but consider the global network structure to determine DEGs in each single condition. And because of using information outside array, it does not need any assumptions on the distributions of expression profiles. We also investigate the validity of our basic assumption that relational information in existing pathway knowledge reflects the expression level, or moreover network structure of microarray data.

Knowledge about pathways covers only small portion of genes that exist (Chung *et al.*, 2004). In case of yeast, less than 10% of the whole genes (i.e. ~6,000) are found to participate in some known pathways. And graph structure is too simple to use as a template structure to compare the pathway and microarray data, because it ignore the orders and types of biological relations. These incompleteness and simplicity may give insignificant results of our model. But applying other biological information, such as protein interaction data and relational information of genes or proteins from biological literature, and modifying graph structure into more complicated structure, we can overcome the problem of our current model.

Acknowledgements

This study was supported by a grant from Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (0412-MI01-0416-0002).

References

Chartrand, G., Kubicki, G., and Schultz, M. (1998). Graph similarity and distance in graphs. *Aequationes Math.* 55, 129-145.

Chung, H.J., Kim, M., Park, C.H., Kim, J., and Kim, J.H. (2004). ArrayXPath: mapping and visualizing microarray gene expression data with integrated biological pathway resources using Scalable Vector Graphics. *Nucleic Acids Res.* 32, W464-W464.

DeRisi, J.L., Iyer, V.R., and Brown, P.O. (1997). Exploring the metabolic and genetic control of gene expression as a genomic scale. *Science* 278, 680-686.

Dudoit, S., Yang, Y.H., Gallow, M.J., and Speed, T.P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 12, 111-139.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* 23, 70-86.

Farkas, I., Jeong, H., Vicsek, T., Barabási, A.-L., and Oltvai, Z.N. (2003). The topology of the transcription regulatory network in the yeast, Saccharomyces cerevisiae. *Physica* A 318, 601-612.

Holme, P., Huss, M., and Jeong, H. (2003). Subnetwork hierarchies of biochemical pathways. *Bioinformatics* 19, 532-538.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemarker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., and Friend, S.H. (2000). Functional Discovery via a Compendium of expression Profiles. *Cell* 102, 109-126.

Hunter, L., Taylor, R.C., Leach, S.M., and Simon, R. (2001). GEST: a gene expression search tool based on a novel Bayesian similarity metric. *Bioinformatics* 17, Suppl.1, S115-S122.

Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C., Trent, J.M., Staudt, L.M., Hudson, J. Jr, Boguski, M.S., Lashkari, D., Shalon, D., Botstein, D., and Brown P.O. (1999). The transcriptional program in the response of human fibroblasts to serum. *Science* 83-87, 283.

Kanehisa, M. (1996). Toward pathway engineering: a new database of genetic and molecular pathways. *Science & Technology Japan* No. 59, pp. 34-38.

Kanehisa, M. (2000). Post-Genome Informatics. Oxford University Press.

Kharcheko, P., Vitkuo, D., and Church, G.M. (2004). Filling gaps in a metabolic networks using expression information. *Bioinformatics* 20, Suppl. 1, I178-I185.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., and Tsui, K.W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression change from microarray data. *Journal of Computational Biology* 8, 37-52.

Qian, J., Lin, J., Luscombe, N.M., Yu, H., and Gerstein, M. (2003). Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics* 19, 1917-1926.

Rung, J., Schlitt, T., Brazma, A., Freivalds, K., and Vilo, J. (2002). Building and analysing genome-wide gene disruption networks. *Bioinformatics* 18, Suppl. 2, S202-S210.

Shmulevich, I. and Zhang, W. (2002). Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics* 18, 555-565.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* 98, 5116- 5121.

Wagner, A. (2001). How to reconstruct a large genetic network from n gene perturbations in fewer than n2 easy steps. *Bioinformatics* 17, 1183-1197.

Yamanishi, Y., Vert, J.P., Nakaya, A., and Kanehisa, M. (2004). Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics* 20, Suppl. 1, I363-I370.

Zhao, Y. and Pan, W. (2003). Modified nonparametric approached to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics* 19, 1046-1054.