

# Biological Pathway Extension Using Microarray Gene Expression Data

Tae Su Chung<sup>1,2†</sup>, Jihun Kim<sup>1†</sup>, Keewon Kim<sup>1</sup>  
and Ju Han Kim<sup>1,2\*</sup>

<sup>1</sup>Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, Seoul 110-799, Korea, <sup>2</sup>Human Genome Research Institute, Seoul National University College of Medicine, Seoul 110-799, Korea

## Abstract

Biological pathways are known as collections of knowledge of certain biological processes. Although knowledge about a pathway is quite significant to further analysis, it covers only tiny portion of genes that exists. In this paper, we suggest a model to extend each individual pathway using a microarray expression data based on the known knowledge about the pathway. We take the Rosetta compendium dataset to extend pathways of *Saccharomyces cerevisiae* obtained from KEGG (Kyoto Encyclopedia of genes and genomes) database. Before applying our model, we verify the underlying assumption that microarray data reflect the interactive knowledge from pathway, and we evaluate our scoring system by introducing performance function. In the last step, we validate proposed candidates with the help of another type of biological information. We introduced a pathway extending model using its intrinsic structure and microarray expression data. The model provides the suitable candidate genes for each single biological pathway to extend it.

**Keywords:** biological pathway, pathway extension, microarray gene expression

## Introduction

Biological pathways can be regarded as a collection of known relations or reactions between biological objects (*i.e.* genes or gene products). But the knowledge about biological pathway is not complete and insufficient to cover the whole number of genes or gene products. In case of human, for example, about 3,000 genes are

covered by the major biological pathway databases. The rest with the relation with other genes remains to be explored. We view biological pathways as a set of biological knowledge that should be expanded by using experimental data like DNA microarray gene expression data.

Recently, microarray gene expression datasets are incessantly cumulated with the aid of recent technological advances. Analyzing these datasets makes it possible to measure the expressions of tens of thousands of genes simultaneously under various experimental conditions. In this paper, we propose a model, PathPlus, to determine genes for biological pathway extension by combining the knowledge about biological pathways and the experimental data from DNA microarray experiments.

The most significant feature of PathPlus is that it uses established knowledge (*i.e.* biological pathway) as a template to expand the knowledge. Previous studies try to directly infer biological networks (or pathways) from gene expression data by various computational or statistical means without using prior knowledge (or template). To induce networks of genes or proteins, for example, that might conjecture biological pathways from large-scale data, some construct bayesian networks from gene expression data or infer functional protein complexes from protein interaction data (Friedman *et al.*, 2000). Others combine various genomic data to infer biological networks without using prior knowledge about biological interaction. In fact, there are a few recent works trying to reconstruct biological relations based on prior knowledge (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 approach. The work of Kharchenko *et al.* compares established metabolic network with expression profiles to find genes that can complete a metabolic pathway with some participants missed. While the methods are good in finding missing genes, they do not suggest possible new members (or genes) for the given biological pathway for pathway extension.

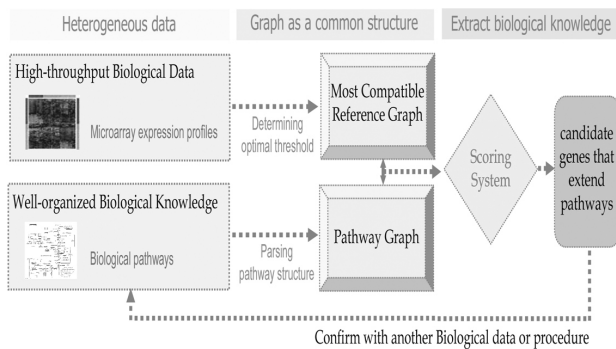
We first observe that a biological pathway contains highly verified information but covers only a small fraction of genes, while microarray data provide noisy experimental data but covers the whole genome. The essence of PathPlus approach is to determine candidate genes that are highly likely to be related to a given pathway by combining microarray gene expression data

\*Corresponding author: E-mail juhan@snu.ac.kr

†Contributed equally

Tel +82-2-740-8320, Fax +82-2-762-3063

Accepted 19 October 2008



**Fig. 1.** Overview of PathPlus. To combine biological pathway and microarray gene expression data, that have different data structures, we transform both into graphs, pathway graphs and reference graphs. Applying a scoring system, we obtain a set of candidate genes with high scores for each pathway. The set is verified by using another knowledge resource, the Gene Ontology.

and the known knowledge about the biological pathway (Fig. 1).

To combine biological pathways and microarray datasets, that are very different in data structure, we transform both into a common template structure, a graph. We transform each pathway into a graph whose nodes correspond to genes or gene products and edges to biological relations between them. Then, we extract from microarray data a reference graph having the largest structural similarity with the given pathway graph by optimizing the parameters of graph construction. PathPlus provides a score measuring the likelihood of a gene to be related to a pathway using the structure of the reference graph. The scoring system is based on the centrality measures introduced by the social network analysis (Bogatti *et al.*, 1997; Faust, 1997).

Finally, we obtain sets of candidate genes with high scores for biological pathways and evaluate the results by verifying the functions of the candidate genes using another heterogeneous data, the Gene Ontology (GO), with promising results. PathPlus approach may contribute to establish methodological concepts in knowledge extension and to give insights into biologically untapped territories.

## Methods

To combine the two different data types, biological pathways and microarray gene-expression data, PathPlus use graph as the common template for analysis. A graph consists of a set of nodes and a set of edges. Nodes represent genes (or gene products) and edges gene-to-gene interactions in the present study.

### Graph construction from biological pathway

For the purpose of illustration in the present study, we use pathways and gene expression profiles of *Saccharomyces cerevisiae*. KEGG (Kyoto Encyclopedia of genes and genomes) database (Kanehisa, 1996) provides 88 biological pathways for *Saccharomyces cerevisiae*. We extracted 43 pathways by removing those containing less than 12 genes to avoid the perturbation caused by the scarcity of basis knowledge. KEGG relations are divided into three categories: EC relation, PP relations and GE relation. EC relation stands for relations between two genes whose protein products share the same metabolite in a metabolic pathway. When two proteins interact directly, genes coding them are said to have PP relation. GE relation means that one gene or its product regulates the expression of the other gene. In addition, we define co-member relation that refers to the genes assigned to the same function in a pathway. To construct a pathway graph from this information, we make a node for each gene and link a pair of nodes when they are assigned one of the relations listed above.

### Graph construction from microarray expression data

As a source of microarray expression data, we use the Rosetta compendium dataset (Hughes *et al.*, 2000), which is hitherto the most systematic approach to profile yeast genes. 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.

The whole genes in the microarray dataset make nodes in reference graph of the pathway graph for each pathway we will concern. We here present two different methods to construct graph from the microarray dataset. Each method needs thresholds in constructing graphs and we will give criteria to optimize threshold so that the resulting graph reflects the biological knowledge in some sense.

**Method of linking co-DEGs:** The process of identifying DEGs (differentially expressed genes) is one of the major steps of analyzing gene expression data under the assumption that the expression data are modeled to find significant changes in gene expression levels between treatment conditions and control conditions.

Many reasonable procedures and statistical methods

for optimizing threshold to determining DEGs are introduced and applied to further analysis (Iyer *et al.*, 1999; Tusher *et al.*, 2001; Dudoit *et al.*, 2002). Linking co-DEGs methods constructs a graph by linking two genes that are co-DEGs, *i.e.*, they are DEGs under same experimental condition. The method was applied to the microarray expression dataset with various threshold values.

**Method of linking similarly expressed genes:** This method is the way of determining edges based on expression similarity. In this paper, the absolute value of Pearson correlation coefficient is taken as similarity system. When a similarity of a pair of genes exceeds a given threshold, these two genes are linked in this method.

**The most compatible reference graph of pathway graph**

Among all possible values of threshold, single value is to be determined so that the compatibility between pathway graph and reference graph can be maximized. The compatibility between two graph structures is designed to indicate the measurement by which two graph structures are similar. Here we will introduce the notion of compatibility between two graphs, in the general context of graph theory. The geodesic distance  $d(g, h; G)$  between nodes  $g$  and  $h$  is defined by the length of shortest path from  $g$  to  $h$  in the graph  $G$ . This distance rep-

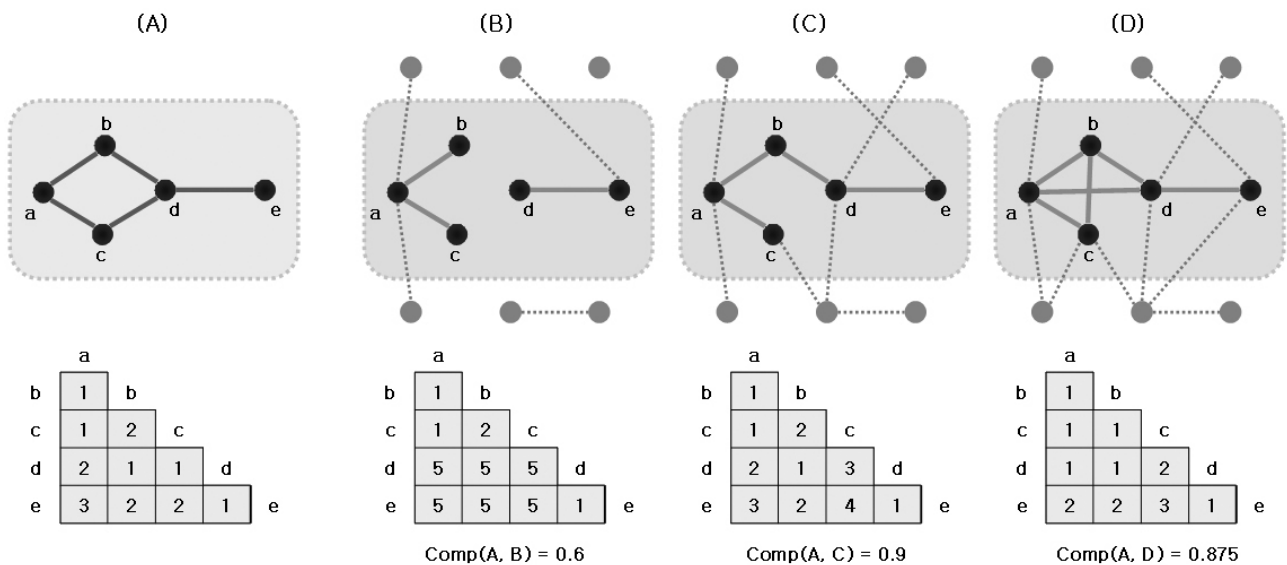
resents the global structure of graph (Chartrand *et al.*, 1988). If two graphs  $G_1$  and  $G_2$  are constructed on the same set of nodes, then the geodesic distance of two graphs can be easily defined by the average of differences of all geodesic distances of all pairs of nodes in each graph, *i.e.*

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

where the summation is taken over all (unordered) pairs of nodes and  $n$  is the number of common set of nodes. We here note that it is symmetric and satisfies triangle inequality (Chartrand *et al.*, 1988). Now let  $G_P$  be the pathway graph and  $G_M = G_M(q)$  is a (reference) graph constructed by microarray data with threshold  $q$  using certain graph construction method. Then the compatibility  $Comp(G_P, G_M)$  is obtained by

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

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 small 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  $G_M$  includes exactly same suture of  $G_P$ . The most compatible reference graph of  $G_P$  is the graph  $G_M$  which maximize the compatibility with  $G_P$ . Simple greedy algorithm is used to find such optimal threshold for each pathway. It is



**Fig. 2.** Compatibilities between pathway graph and various reference graphs. When a pathway graph is given at (A), various reference graphs can be compared to it according to different parameters ((B), (C) and (D)). Triangular tables represent geodesic distances of gene pairs in pathway graph (A) and reference graphs (B)-(D) which are restricted to genes in the pathway, respectively. In this figure, (C) is determined to be the most compatible reference graph of (A).

worthy to note that different optimal threshold may used for extending different pathways.

The effect of threshold optimization on compatibility is illustrated in Fig. 2. When a pathway graph is given in (A), various reference graphs can be compared to it according to different parameters ((B), (C) and (D)). In this figure, (C) is determined to be the most compatible reference graph of (A).

**Scoring system**

With the most compatible reference graph of a pathway graph, we suggest a gene list supposed to be related to the given pathway by investigating scoring system which is designed to assign a higher score on a gene that is more likely to be related to the pathway. We invent scoring system based on the idea of centrality measures introduced by social network analysis (Bogatti *et al.*, 1997). Social network analysis has its strength in reflecting the structure of a graph being observed, though it does not directly represent quantitative relations between two nodes. Because many biological networks are known to work in on-off manner, we suggest that social network analysis would be a suitable method to analyze biological networks. Here in this paper, we make our scoring system by adopting closeness centrality and betweenness centrality from social network analysis (Bogatti *et al.*, 1997; Faust, 1997). The normalized betweenness centrality  $C_{bet}(g)$  and normalized closeness centrality  $C_{clo}(g)$  of a node  $g$  are defined by

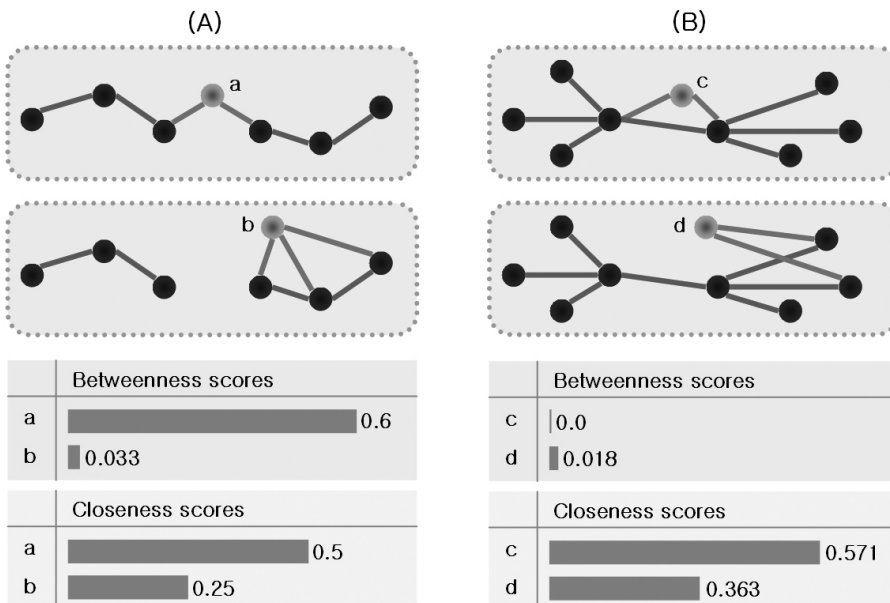
$$C_{bet}(g) = \frac{1}{(n-1)(n-2)/2} \sum_{\substack{x \neq g \\ y \neq g}} \frac{path(x,y;g)}{path(x,y)}$$

and

$$C_{clo}(g) = \frac{n-1}{\sum_{x \neq g} d(g,x)}$$

Here  $n$  is the number of nodes in the graph,  $path(x, y)$  is the number of all shortest paths from  $x$  to  $y$ , and  $path(x, y; g)$  is the number of shortest paths from  $x$  and  $y$ , which pass through  $g$ . In the definition of closeness centrality, the geodesic distance  $d(g, x)$  is used. Betweenness centrality of a node measures the potential to mediate flows of resources or information between other nodes, and closeness centrality of a node measures the distances by efficient paths from the node to all the other nodes.

Fig. 3 illustrates the different behaviors of the two centrality measures in a certain types of graphs. In (A), centrality of  $a$  is expected to be greater than that of  $b$ , since the node  $a$  make connection of two disconnected subgraphs, while the node  $b$  is related with only single subgraph. Although both centralities of  $a$  are greater than those of  $b$ , betweenness centralities of  $a$  and  $b$  distinguish their social roles dramatically. On the other hand, in (B), closeness centralities can discriminate the social roles rather than betweenness centralities. In this case, related nodes with  $c$  are more central than nodes related with  $d$ . To assign a score  $S(g)$  of a gene  $g$ , we need to construct a temporary graph which is a subgraph of the most compatible reference graph  $G_M$ , whose nodes are set of genes in the pathway graph  $G_P$



**Fig. 3.** Distribution of scores of specific nodes in certain types of graphs. In (A), betweenness scores of  $a$  and  $b$  distinguish social roles of  $a$  and  $b$  dramatically, while closeness scores discriminates social roles of  $c$  and  $d$  in (B).

plus a gene  $g$ . We define two scoring systems by normalized betweenness and closeness centrality measures of  $g$  in the temporary graph. We call them the betweenness scoring system and closeness scoring system, respectively.

## Results and Discussion

### Correlation between microarray data and pathway knowledge

The basic assumption of PathPlus is that gene expression profile reflects pathway structure and pathway structure affects gene expression profile. We defined two ‘expressional homogeneity’ indices, co-degree and similarity, between a pair of genes.

Co-degree between two genes in a microarray gene-expression dataset is defined as the number of conditions in which both are differentially expressed. Similarity between two genes is defined as the absolute value of the Pearson correlation between the two gene profile vectors. Two genes directly connected in a pathway graph are expected to have higher homogeneity indices than those non-connected.

We further defined three more measures,  $E(P,M)$  = the average of the homogeneity (*i.e.* co-degree or similarity) indices of all directly connected pairs in a pathway graph  $P$ ,  $N(P,M)$  = the average of the homogeneity (*i.e.* co-degree or similarity) indices of all pairs that are

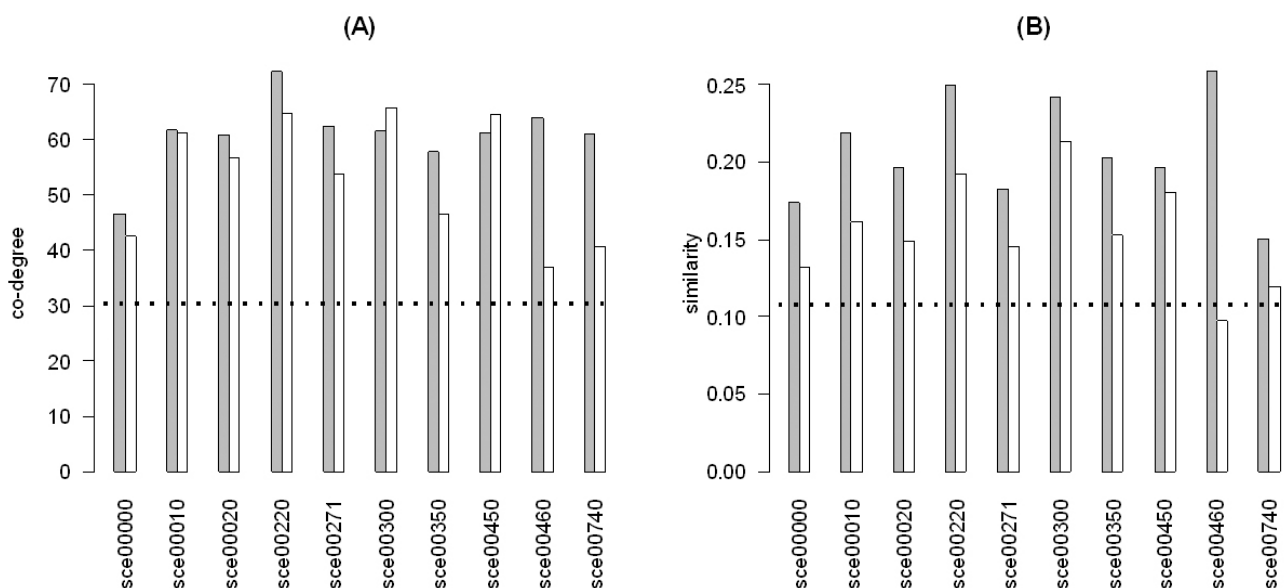
not directly connected in a pathway graph  $P$ , and  $R(M)$  = the average of homogeneity indices of all gene pairs in a microarray dataset  $M$ .

If the basic assumption of PathPlus holds, we expect that the inequality  $E(P,M) > N(P,M) > R(M)$  should be satisfied. Fig. 4 demonstrates the distributions of the average homogeneity indices of directly connected and non-directly-connected gene pairs for selected pathways and those of all gene pairs from the whole microarray dataset. Because it is observed that the inequality holds, it is suggested that microarray gene-expression profile reflects biological pathway structures and vice versa. These measures were only introduced to show the correlation between microarray data and pathway knowledge.

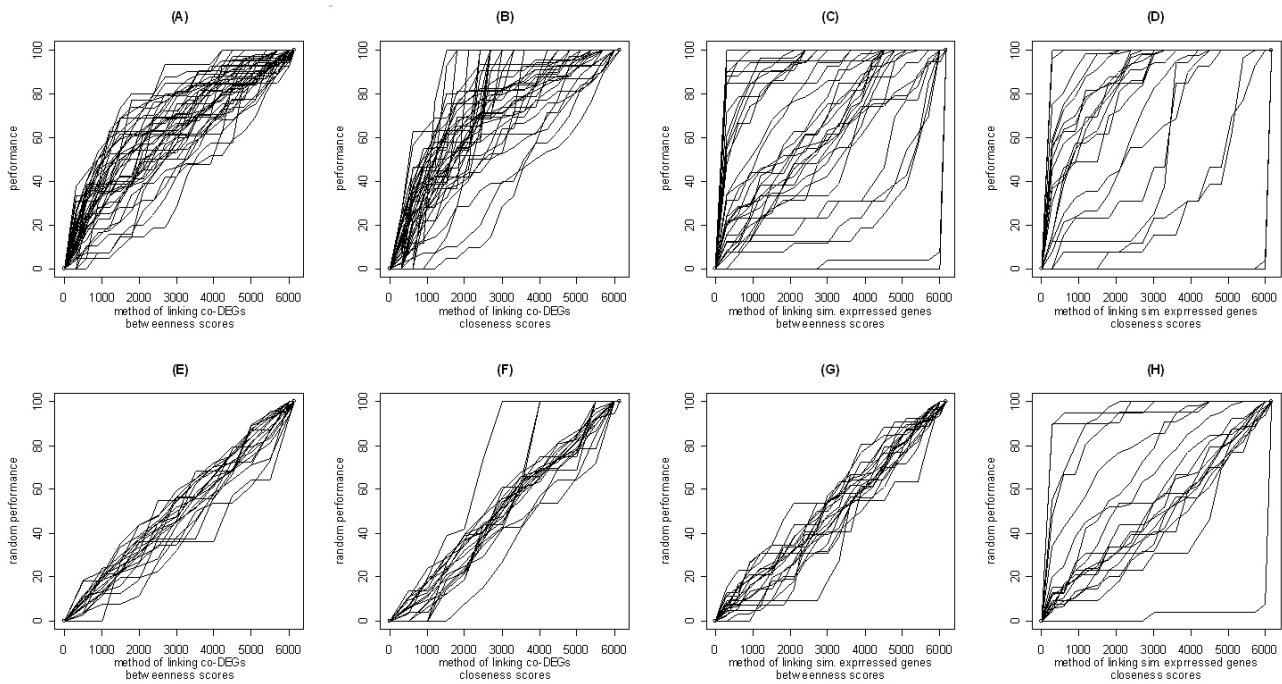
### Performance evaluation of scoring systems

PathPlus assigns a score for each candidate gene. We evaluated the reliability of the scoring system. To measure the reliability of scoring system or performance of model, we introduce a performance function,  $f(k) = M(k)/n \cdot 100$  (%), where  $M(k)$  is the number of genes that included in the original pathway among top  $k$  ranked genes, and  $n$  is the number of genes in the pathway, and  $k$  runs from 0 to  $N$ , the number of candidate genes. ( $N=6,153$  in this study).

Fig. 5 shows performance functions of each scoring systems in each graph-construction methods. It also



**Fig. 4.** Correlation between microarray data and knowledge about selected pathways. Both (A) Co-degree and (B) similarity of connected gene pairs (*i.e.* the gray bars) show higher values than non-connected ones (*i.e.* the white bars) for random selected 10 pathways. Overall results are similar to this result. The dotted horizontal lines represent the mean homogeneity values of the whole gene pairs.



**Fig. 5.** Performance functions of pathways. (A), (B), (E), (F) use reference graph constructed by linking co-DEGs, (C), (D), (G), (H) use method of linking similarly expressed genes. The first and third columns applied by betweenness scoring system, and the remains are applied by closeness one. The upper part give performances of actual pathways and the lower part show performances of randomly constructed pathways.

provided the performance of randomly constructed pathways.

Fig. 5 (A), (B), (E) and (F) are performance functions of pathways using the reference graph constructed by the method of linking co-DEGs, while (C), (D), (G) and (H) are using the method of linking similarly expressed genes. The (A), (C), (E) and (G) are applied by the betweenness scoring system, and the remaining part (B), (D), (F) and (H) are applied by the closeness scoring system. The upper part (A), (B), (C) and (D) show performances of actual pathways, but the lower part (E), (F), (G) and (H) represent performances of randomly constructed pathways. As expected, performances functions of randomly constructed pathways are roughly located on the diagonal lines connecting origin and top-right point ( $N, 100$ ). The curves of actual performances are located in the upper triangular parts in general, so we expect that two scoring systems make quite good results of our purpose. Betweenness scoring systems give more stable performance with the reference graph constructed by linking co-DEGs. The reference graph obtained by the method of linking similarly expressed genes serves not-so-good performance scoring systems. These poor results maybe related to the similarity system (the absolute value of Pearson correlation coefficient) of two genes. Pearson correlation usually ig-

nores the local similarity, *i.e.*, it may assign low similarity to pair of genes which are highly co-expressed only in the small number of experimental conditions. The linking co-DEGs methods overcome the locality problem in some sense.

### Validation of the model via extrinsic information

The candidate gene list has been obtained by applying our model to each pathway. Since we have two scoring systems and also two methods to construct reference graphs, each pathway has 4 lists of candidates, Table 1 and 2 show the top-ranked genes with respect to the pathways: "glycolysis / gluconeogenesis" and "TCA cycle". These two pathways are selected due to their biological significances and performances. In these tables, we find that ranks of gene list from different scoring systems are quite compatible each other, but different types of constructing reference graph create inconsistent lists of candidates. Each table consists of two lists classified by its methods of graph construction, and in each part, sums of ranks ( $r$ ), ranks in betweenness scoring system ( $rb$ ), ranks in closeness scoring system ( $rc$ ) and grades are examined. The grades are assigned by the association with amiGO annotation and standard biological references. The grade "A" is assigned to

**Table 1.** Candidate genes to extend pathway: Glycolysis/ Gluconeogenesis

Linking co-DEGs					Linking similarly expressed genes				
ORF	r	rb	rc	Grade	ORF	r	rb	rc	Grade
YFL014W	3	2	1	B	YBR019C	2	1	1	A
YGL255W	5	3	2	C	YMR323W	5	3	2	B
YML128C	7	4	3	?	YLR377C	6	2	4	A
YBR296C	11	7	4	?	YKR097W	7	4	3	A
YGL117W	13	8	5	?	YPL017C	11	5	6	B
YLR303W	13	5	8	B	YGL256W	12	7	5	B
YFR053C	14	1	13	A	YAL054C	13	6	7	A
YHR136C	15	9	6	B	YDR050C	18	10	8	A
YKL096W	17	10	7	B	YDR045C	19	9	10	C
YDL037C	23	6	17	?	YJR141W	19	8	11	?
YCL018W	27	11	16	B	YDR380W	20	11	9	A
YBR147W	29	14	15	?	YPL146C	25	13	12	?
YGR110W	33	12	21	?	YNL001W	26	12	14	?
YML123C	33	24	9	A	YMR233W	34	14	20	?
YGR035C	35	15	20	?	YJL068C	37	20	17	B
YMR094W	35	25	10	C	YCR095C	40	17	23	?
YMR096W	37	26	11	C	YLR336C	42	23	19	?
YNL160W	39	27	12	A	YJR003C	48	30	18	?
YNR058W	48	13	35	C	YLR287C	50	21	29	?
YBR047W	50	36	14	?	YPR169W	50	18	32	?

In the table, r is for sums of ranks, rb for betweenness ranks, rc for closeness ranks. The grades are assigned by the association with amiGO annotation. The grade “A” is for genes that clearly associated with the pathway, “B” is for genes that have equivocal interpretation, “C” is for genes without association and “?” is for genes with unknown functional annotation.

**Table 2.** Candidate genes to extend pathway: TCA cycle

Linking co-DEGs					Linking similarly expressed genes				
ORF	r	rb	rc	Grade	ORF	r	rb	rc	Grade
YBR296C	2	1	1	A	YJL045W	2	1	1	A
YFL014W	4	2	2	B	YKR097W	5	2	3	A
YGL255W	6	3	3	C	YMR323W	5	3	2	A
YML128C	8	4	4	?	YDR380W	9	4	5	A
YIR039C	12	5	7	B	YDR148C	10	6	4	?
YGL117W	15	9	6	?	YPL017C	11	5	6	B
YBR072W	17	7	10	B	YHR196W	18	8	10	C
YLR303W	19	10	9	C	YDR299W	24	10	14	?
YCL040W	21	16	5	A	YPL146C	24	13	11	?
YJL159W	25	17	8	C	YGR128C	26	11	15	B
YOR383C	33	8	25	?	YJR041C	26	7	19	B
YPR167C	34	6	28	C	YJR003C	27	9	18	?
YFR053C	35	22	13	A	YLR051C	33	12	21	?
YKL096W	38	21	17	C	YJL068C	38	21	17	B
YJL116C	39	24	15	?	YLR022C	38	18	20	?
YCL018W	45	12	33	C	YPL210C	40	16	24	C
YJL079C	47	33	14	?	YBR142W	43	30	13	B
YMR276W	47	26	21	C	YLR435W	43	20	23	?
YOR382W	47	23	24	?	YHR197W	45	14	31	?
YJR028W	50	34	16	C	YBR242W	50	23	27	?

In the table, r is for sums of ranks, rb for betweenness ranks, rc for closeness ranks. The grades are assigned by the association with amiGO annotation. The grade “A” is for genes that clearly associated with the pathway, “B” is for genes that have equivocal interpretation, “C” is for genes without association and “?” is for genes with unknown functional annotation.

genes that are clearly associated with the pathway, "B" is genes with equivocal interpretation, "C" is genes without association and "?" is for the case of unknown functional annotation. All grade is endowed with subjectivity.

Let us look into the candidates for "glycolysis / gluconeogenesis" pathway. A few candidates are annotated a general or vital cellular function or component that we cannot easily judge the relevance to the pathway (YFL014W, YML128C, YHR136C, YKL096W and YMR094W). A few others are not revealed about their functions (YBR296C, YGL117W, YDL037C, YGR110W, YBR147W and YGR035C). And some genes suggestive of strong relation with the pathway also have been found (YFR053C, YML123C, YNL160W and YBR296C). YFR053C functions as a hexokinase, which is one of the key enzymes of glycolysis but has been omitted from pathway graph because the gene ID has not been served by KEGG database (E.C 2.7.1.1, in glycolysis / gluconeogenesis pathway data from KEGG corresponds to it). YBR296C and YML123C are genes whose products function as phosphate transporters-glycolysis process is intrinsically linked with citric acid cycle, which is subsequently linked with oxidative phosphorylation. Therefore the extension of glycolysis / gluconeogenesis pathway toward oxidative phosphorylation via phosphate transporter can be sufficiently persuasive.

The next candidate list is the one for extension of citric acid cycle. This pathway is historically well known, and is considered to be centrally located within a metabolic pathway network. (Jeong *et al.*, 2001) Hence almost every gene might be traced to its relation to citric acid cycle within a few connections.

Nevertheless, in the list are some genes that draw our attention (YLR303W, YCL040W, YCL018W, YOL151W, YCL030C, YDL022W, YFR053C and YJL116C). YLR303W deals with serine metabolism, which is initiated from a precursor of citric acid cycle, 3-phosphoglycerate. YCL040W is a gene with glucokinase activity that is essential in glycolysis/glucokinase pathway, which, as we mentioned above, is closely lined with citric acid cycle. YJL116C is a gene for mitochondrion organization and biogenesis and citric acid cycle occurs within mitochondrion. YOL151W is known to have oxidoreductase activity, which is crucial for oxidative phosphorylation- the very next step of citric acid cycle for energy generation. YDL022W is related to NAD activity, which plays a key role in citric acid cycle. Similarly, others participate in or are related to citric acid cycle; YCL018W is related to the cycle through malate metabolism, YCL020C through histidine metabolism and YFR053C through fructose metabolism.

## Conclusions

In this paper, we made an attempt to extend an established pathway through microarray expression data. To obtain this goal, we transformed knowledge about pathways and microarray expression data, respectively, into graph structures, and analyzed two graphs with a scoring system we invented based on social network theory. We also validated the reliability of the scoring system by computing prediction powers. Then final result of candidate lists has been presented and has been examined through another source of biological data, amiGO annotation.

## References

- Bogatti, S.P., and Everett, M.G. (1997). Network analysis of 2-mode data. *Social Networks* 19, 243-269.
- Chartrand, G., Kubicki, G., and Schultz, M. (1988). Graph similarity and distance in graphs. *Aequationes Math.* 55, 129-145.
- 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.
- Faust, K. (1997). Centrality in affiliation networks. *Social Networks* 19, 157-191.
- Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian Network to Analyze Expression Data. *Journal of Computational Biology* 7, 601-620.
- 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., Shoemaker, D.D., Gachotte, D., Chakrabarty, K., Simon, J., Bard, M., and Friend, S.H. (2000). Functional discovery via a compendium of expression profiles. *Cell* 102, 109-126.
- 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* 283, 83-87.
- Jeong, H., Mason, S.P., Barabasi, A.L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. *Nature* 409, 41-2.
- Kanehisa, M. (1996). Toward pathway engineering: a new database of genetic and molecular pathways. *Science & Technology Japan* 59, 34-38.
- Kharchenko, P., Vitkuo, D., and Church, G.M. (2004). Filling gaps in a metabolic networks using expression information. *Bioinformatics* 20 (Suppl. 1), i178-i185.
- 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.
- Yamanishi, Y., Vert, J.P., Nakaya, A., and Kanehisa, M. (2004). Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics* 20 (Supple. 1), i363-i370.