Extraction of Correlated Gene Clusters by Multiple Graph Comparison

Akihiro Nakaya nakaya@kuicr.kyoto-u.ac.jp
Susumu Goto goto@kuicr.kyoto-u.ac.jp
Minoru Kanehisa kanehisa@kuicr.kyoto-u.ac.jp
Bioinformatics Center, Institute for Chemical Research, Kyoto University
Uji, Kyoto 611-0011, Japan

Abstract

This paper presents a new method to extract a set of correlated genes with respect to multiple biological features. Relationships among genes on a specific feature are encoded as a graph structure whose nodes correspond to genes. For example, the genome is a graph representing positional correlations of genes on the chromosome, the pathway is a graph representing functional correlations of gene products, and the expression profile is a graph representing gene expression similarities. When a set of genes are localized in a single graph, such as a gene cluster on the chromosome, an enzyme cluster in the metabolic pathway, or a set of coexpressed genes in the microarray gene expression profile, this may suggest a functional link among those genes. The functional link would become stronger when the clusters are correlated; namely, when a set of corresponding genes form clusters in multiple graphs. The newly introduced heuristic algorithm extracts such *correlated gene clusters* as isomorphic subgraphs in multiple graphs by using inter-graph links that are defined based on biological relevance. Using the method, we found *E.coli* correlated gene clusters in which genes are related with respect to the positions in the genome and the metabolic pathway, as well as the 3D structural similarity. We also analyzed protein-protein interaction data by two-hybrid experiments and gene coexpression data by microarrays in *S.cerevisiae*, and estimated the possibility of utilizing our method for screening the datasets that are likely to contain many false positive relations.

Keywords: correlated gene cluster, binary relationships, graph comparison, clustering

1 Introduction

Correlated gene clusters. The complete genome sequence contains the information about ordering of genes along the chromosome. Besides such *geometrical* relationships, other features characterize relationships among genes, including *similarity* relationships based on sequences or 3D structures of gene products, and *functional* relationships in metabolic/regulatory pathways. When multiple gene-gene relationships can be found on different attributes as above, it would be interesting to see whether or not a set of genes share their mutual relationships in relation to each attribute. For example, as reviewed by Erlandsen *et al.* [2], the enzymes in the glycolytic pathway (Fig. 1) commonly display α/β folds, which is obtained by examining the relationships of enzymes with respect to their structural similarities and neighboring relationships in the pathway. This type of observation has been made for a specific set of genes. Here we examine all the sets of genes in a given organism that are correlated with respect to more than one attribute.

Gene-gene relationships on a specific attribute can be denoted by using a set of *binary relationships* in a general manner. For example, let a binary operator '$\sim$' denote a binary relationship between two genes, and let g_1, g_2, g_3, and g_4 be a series of genes arranged in this order in a genome sequence, their geometrical relationships are broken down into a set of binary relationships $\{g_1 \sim g_2, g_2 \sim g_3, g_3 \sim g_4\}$. A set of binary relationships among genes forms a graph structure as a whole. Fig. 2 shows three graphs G_1 (genome), G_2 (pathway), and G_3 (similarity), where each graph node corresponds to a

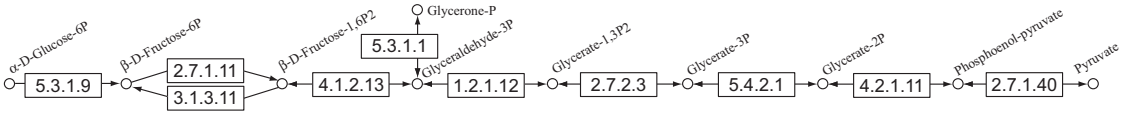

Figure 1: A series of enzymes in the glycolytic pathway display α/β folds.

gene or a gene product. In a graph, two nodes are connected by an edge (expressed by a solid line) when they are related by a binary relationship. In a set of genes, if all or most of the genes reserve their mutual relationships in multiple graphs, like the light gray nodes and the dark gray nodes in Fig. 2, the biological relevance among those genes is considered to be supported at high possibility. We call such a set of genes a *correlated gene cluster* (or simply, *correlated cluster*), by which we can characterize, classify, and predict the activities of genes.

Hyperedges — Introduction of inter-graph links. Finding correlated gene clusters can be formalized as a subgraph isomorphism problem and has been proved to be NP-complete. Therefore, some heuristics are required to cope with this problem. To extract correlated gene clusters from two graphs, Ogata *et al.* [5] introduced a notion called FRECs (Functionally Related Enzyme Clusters). They also introduced a set of inter-graph links between two nodes that correspond to the same gene in each graph, and searched isomorphic subgraphs in the two graphs so that the nodes of the subgraphs are connected by the inter-graph links. By comparing the genome and the metabolic pathway, they found that seven *Escherichia coli* genes catalyzing successive reaction steps in peptidoglycan biosynthesis pathway are located in close position also along the genome sequence, for instance.

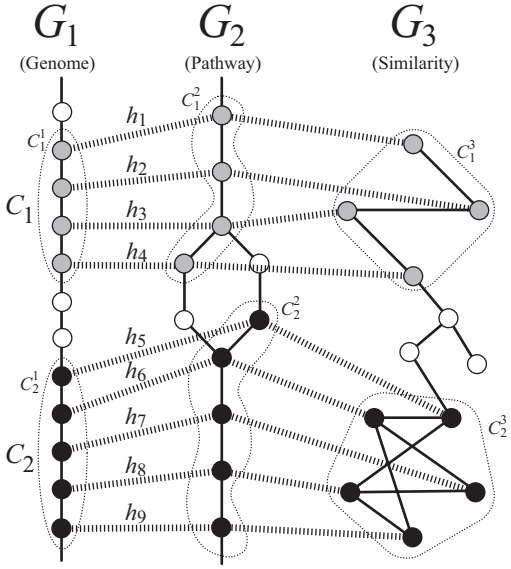

Figure 2: Correlated gene clusters.

We can extend the notion of FRECs by increasing the number of graphs so that the additional graphs provide information about gene-gene relations that cannot be found by just two graphs. Fig. 2 shows an example of correlated gene clusters (C_1 and C_2) in three graphs G_1, G_2, and G_3. Here, dashed lines represent linking of corresponding genes from three graphs and grouping them into a single category (same gene). We call this linkage a *hyperedge* and define a distance between two hyperedges so that it can reflect the distance (typically, the shortest path length) between their nodes in each graph. By gathering hyperedges based on this distance, we can find a set of nodes that are tightly coupled in the graphs, that is, a correlated gene cluster.

Correlation among known and unknown genes. Recent high-throughput experimental technologies provide huge and potentially interesting datasets, but they often contain unknown and hypothetical relationships as well as erroneous relationships. For example, coexpression relationships by microarrays/oligochips and protein-protein interactions by two-hybrid analysis might have such characteristics. Today's standard approaches to analyze such datasets include, for example, clustering the genes according to similarity of expression patterns, and extracting densely connected network components in a protein-protein network. When a set of known genes can be placed into the same category, such approaches may uncover functional links to unknown genes based on some biological features. Our method basically follows this strategy. One of the advantages of our method over the existing ones is that it makes it possible to automatically incorporate multiple features observed in multiple datasets. Even if relationships among genes cannot be explained in a single graph, it is possible to improve the sensitivity of data analysis by evaluating corresponding relationships in additional

graphs.

As a practical application of our method, we will present the correlated gene clusters found in *Escherichia coli* datasets and *Saccharomyces cerevisiae* datasets. The latter includes the two-hybrid protein-protein interaction dataset by Ito *et al.* [3] and the coexpression dataset based on microarray experiments by DeRisi *et al.* [1].

2 Method

Input datasets. As input datasets, we use a set of n graphs $G = \{G_1, \ldots, G_n\}$ and a set of m hyperedges $H = \{h_1, \ldots, h_m\}$. When n graphs are used, we denote a hyperedge with an n-tuple $h_i = (x_{1,i_1}, \ldots, x_{n,i_n})$. Here, the kth element $h_i^k = x_{k,i_k}$ is G_k's node that constitutes the hyperedge $(1 \leq k \leq n)$, and we assume that a hyperedge consists of exactly n nodes to make the problem simple.

Distance between hyperedges. Suppose that there are n graphs. Let $C_1 = \{h_{s_1}, \ldots, h_{s_p}\}$ and $C_2 = \{h_{t_1}, \ldots, h_{t_q}\}$ be sets of hyperedges, and let $C_1^k = \{h_{s_1}^k, \ldots, h_{s_p}^k\}$ and $C_2^k = \{h_{t_1}^k, \ldots, h_{t_q}^k\}$ be sets of the kth elements of hyperedges in C_1 and C_2, respectively. We define the distance between two sets of hyperedges C_1 and C_2 as follows:

$$D(C_1, C_2) = \sum_{1 \leq s \leq n} dis(C_1^s, C_2^s), \tag{1}$$

where $dis(C_1^s, C_2^s)$ is the distance between C_1^s and C_2^s. Here, for example, $dis(C_1^s, C_2^s)$ is defined as $max\{d(x,y) | x \in C_1^s, y \in C_2^s\}$,[1] where $d(x,y)$ is the length of the shortest path between nodes x and y in graph G_s (which can be calculated by Dijkstra's algorithm or Warshall-Floyd's algorithm).

In Fig. 2, let $H = \{h_1, h_2, \ldots, h_9\}$ denote a set of hyperedges, and suppose that they are divided into two distinct sets $C_1 = \{h_1, \ldots, h_4\}$ and $C_2 = \{h_5, \ldots, h_9\}$. Distance between these two sets is $D(C_1, C_2) = \sum_{1 \leq s \leq 3} dis(C_1^s, C_2^s) = dis(C_1^1, C_2^1) + dis(C_1^2, C_2^2) + dis(C_1^3, C_2^3) = 10 + 8 + 8 = 26$.

Clustering of hyperedges. Using the distance D, we cluster the hyperedges. Let C be the initial set of clusters, each of which consists of a single hyperedge, i.e., $C = \{\{h_1\}, \ldots, \{h_m\}\}$. Starting with C, we iterate the procedure to pick two clusters between which the distance is the smallest and to merge them into a new cluster (i.e., hierarchical clustering using the distance D). To avoid distant genes being merged into the same cluster, we use a threshold defined for each graph. Let p_i be the threshold for graph G_i. When the path length between two nodes x and y is greater than p_i in G_i, we change the value of $d(x,y)$ to infinity, and leave the pairs of clusters whose distance is infinity untouched. When there are no cluster pairs whose distance is not infinity we stop the clustering procedure and thus obtain correlated gene clusters. Since two nodes within length p_i can be merged into the same cluster even if they are not directly connected, the parameter p_i makes it possible to find gene clusters that are not strictly conserved in the graphs.

3 Results

We implemented the algorithm by using the C++ language on a part of a SiliconGraphics Origin 3800[2] running under IRIX 6.5. Part of the program (e.g., calculation of N to N shortest paths using Dijkstra's algorithm) is parallelized by the POSIX thread library. We used this system to carry out multiple graph comparison and found correlated gene clusters in *E.coli* and *S.cerevisiae* datasets. Required computing resources depend on datasets. However, to estimate that the memory usage was feasible, we limited it to 512MB by the C shell `limit` command.

[1] This definition carries out "complete linkage clustering". For "single linkage clustering", *min* is used instead of *max*.

[2] 256 MIPS R14000(500MHz) processors, 256GB main memory, and 8MB L2 cache for each processor.

3.1 *E.coli* correlated gene clusters

We first searched correlated gene clusters in the three *E.coli* datasets, whose nodes correspond to *E.coli* genes or gene products. The *E.coli* genome dataset (G_1: 4,396 nodes and 4,396 edges) defines neighboring relationships among genes in the genome sequence. Two genes that are directly next to each other are connected by an edge. The *E.coli* pathway dataset (G_2: 761 nodes and 1,223 edges) defines positional relationships among gene products in the metabolic pathway.[3] The *E.coli* structure similarity dataset (G_3: 538 nodes and 3,823 edges) defines 3D structural similarities among proteins.[4] Two proteins in the same category are connected by an edge. We used a value 1 as the weight of each edge. Besides these three graphs, we used 917 hyperedges connecting genes and their products in the graphs. Note that since a single node can be contained in multiple hyperedges, the number of hyperedges can be greater than that of nodes of a graph.

By applying our algorithm with the threshold parameters $p_1 = 2$, $p_2 = 3$, and $p_3 = 0$ (execution time was 140 seconds), we found correlated gene clusters in the biotin metabolism pathway (Fig. 3(A)) and the tryptophan biosynthesis pathway (Fig. 3(B)). Table 1 shows the list of the *E.coli* genes constituting the correlated gene clusters in those pathways. Those clusters retain mutual relationships of genes with respect to positions in the genome sequence and structural similarity besides the relationships in the pathways. All the genes in the biotin pathway classified as "alpha and beta" (α/β), and the genes in the tryptophan pathway display TIM-barrel structures. We note, however, that these correlations are already known. Although the result confirms the validity of our method for multiple graph comparison, the addition of a third graph was too restrictive to uncover any new findings. Pairwise graph comparison is biologically more meaningful especially when the datasets do not appear to contain erroneous data.

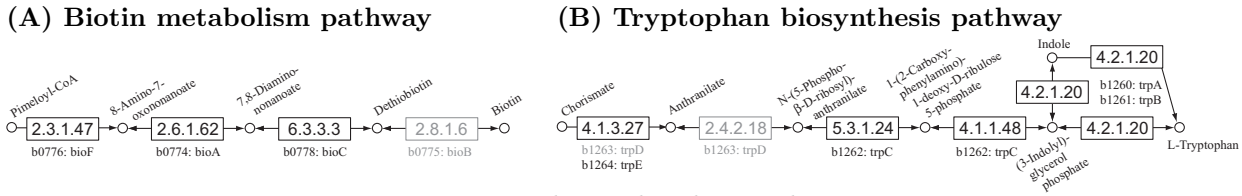

Figure 3: *E.coli* correlated gene clusters.

(A) Biotin metabolism	
b0774	bioA; adenosylmethionine-8-amino-7-oxononanoate aminotransferase (7,8-diamino-pelargonic acid aminotransferase) (dapA aminotransferase) [EC:2.6.1.62] [SP:BIOA_ECOLI]
b0776	bioF; 8-amino-7-oxononanoate synthase (7-keto-8-amino-pelargonic acid synthetase) (7-kap synthetase) (L-alanine–pimelyl CoA ligase) [EC:2.3.1.47] [SP:BIOF_ECOLI]
b0778	bioD; dethiobiotin synthetase (dethiobiotin synthase) (dtb synthetase) (DTBS) [EC:6.3.3.3] [SP:BIOD_ECOLI]
(B) Tryptophan biosynthesis	
b1260	trpA; tryptophan synthase alpha chain [EC:4.2.1.20] [SP:TRPA_ECOLI]
b1261	trpB; tryptophan synthase beta chain [EC:4.2.1.20] [SP:TRPB_ECOLI]
b1262	trpC; indole-3-glycerol phosphate synthase (IGPS) / N-(5'-phospho-ribosyl)anthranilate isomerase (PRAI) [EC:4.1.1.48 5.3.1.24] [SP:TRPC_ECOLI]
b1264	trpE; anthranilate synthase component I [EC:4.1.3.27] [SP:TRPE_ECOLI]

Table 1: *E.coli* genes constituting correlated gene clusters.

3.2 *S.cerevisiae* correlated gene clusters

In this section, we present the results of screening the two-hybrid protein-protein interaction dataset. To evaluate whether or not protein-protein interactions in the dataset are significant, we searched correlated gene clusters. If an interaction or a relation is also observed in biological attributes other than protein-protein interactions, we judge the interaction is relevant.

As the target two-hybrid protein-protein interaction dataset, we used the one developed by Ito *et al.* [3]. This dataset (called "All data" in the original paper), used as the first graph G_1, includes

[3]Compiled by Ogata *et al.* [5]. In this datasets, each node has a unique identifier, but it is related to genes by its EC number that is assigned as a "label", introducing redundancy.

[4]Based on SCOP database release 1.50 [4].
See also "SCOP 3D-fold" from http://www.genome.ad.jp/kegg/kegg2.html.

3,280 genes (nodes) and 4,549 bait-prey interactions (edges). By calculating correlated gene clusters, we compared this dataset with the following datasets (as the second graph G_2) that contain biological relationships: (1) *S.cerevisiae* coexpression dataset, (2) *S.cerevisiae* pathway dataset, and (3) *E.coli* genome dataset.

(1) *S.cerevisiae* two-hybrid v.s. *S.cerevisiae* coexpression

We used the coexpression dataset that was derived from the results of microarray analysis by DeRisi *et al.* [1]. This dataset consists of time-course measurements of gene expressions, and the similarities of chronological changes of expression patterns are compared among gene pairs by means of correlation coefficients.[5] We define a pair of genes are coexpressed when the correlation coefficient is not less than 0.97. Then the dataset consists of 3,307 genes (nodes) and 81,628 coexpressed gene pairs (edges). By using 1,547 hyperedges that connect the identical genes in the two-hybrid dataset and this coexpression dataset, we searched correlated gene clusters (execution time was 48 minutes).

It is interesting to note that these two datasets share only one gene-gene relationship (YGR148C –YLR295C). On the other hand, by including indirect gene-gene relationships through one intermediate gene (i.e., for the ith graph, setting the threshold $p_i = 2$), we found a total of 249 correlated gene clusters. The cluster size ranges from two to nine. Table 2 lists examples of correlated gene clusters that consist of genes whose annotations contain keywords (A) "ribosomal protein", (B) "translation", and (C) "transcription".[6] The genes between two horizontal lines correspond to a correlated gene cluster.

Table 3 shows the correlated gene clusters consisting of the genes that are also referred to in Figure 3 of the original paper by Ito *et al.* [3]. By comparing their two-hybrid dataset with the coexpression dataset by DeRisi *et al.* [1], we found additional sixteen sets of putative gene-gene relationships with respect to (A) autophagy, (B) spindle pole body function, and (C) vesicular transport as shown in Table 3. The genes appear in the paper by Ito *et al.* are hilighted with bold fonts.

(2) *S.cerevisiae* two-hybrid v.s. *S.cerevisiae* pathway

The *S.cerevisiae* pathway dataset (574 nodes and 851 edges) defines positional relationships among gene products in the metabolic pathway.[7] Using 745 hyperedges, we searched correlated gene clusters that reserve their mutual relationships both in the two-hybrid dataset and this pathway dataset (execution time was 32 seconds). The genes that interact through one intermediate gene were included in the same cluster (i.e., for the ith graph, setting $p_i = 2$). Table 4 shows a part of the resulting correlated gene clusters. In this table, we divided the correlated gene clusters into two categories according to whether each cluster contains (A) a single EC number or (B) multiple EC numbers.

In the first category with a single EC number, correlated gene clusters are mainly related to complexes. For example, the first two gene clusters (YJL140W∼YOR210W) and (YJR063W and YDR156W) are related to RNA polymerases. On the other hand, in the category with multiple EC numbers, correlated gene clusters are located in close positions in a pathway. For example, YBR145W and YER073W (being more precise, two enzymes with EC numbers 1.1.1.1 and 1.2.1.3) are next to each other in the bile acids biosynthesis pathway, YFR047C and YLR209C (two enzymes with EC numbers 2.4.2.19, and 2.4.2.1) are next to each other through one intermediate node in the nicotinate and nicotinamide metabolism pathway.

(3) *S.cerevisiae* two-hybrid v.s. *E.coli* genome In the previous parts, we focused on a set of graphs defining the relationships among genes with respect to a single organism (*E.coli* or *S.cerevisiae*), and by using a set of hyperedges connecting the same genes or gene products in multiple graphs, we

[5]See also KEGG/BRITE database at http://www.genome.ad.jp/brite/
and KEGG/EXPRESSION database from http://www.genome.ad.jp/kegg/kegg2.html.

[6]A total list of the correlated gene clusters is obtained from http://web.kuicr.kyoto-u.ac.jp/˜nakaya/pub/giw01/.

[7]Compiled by Ogata *et al.*1 [5]. See also the footnote of section 3.1.

(A) Ribosomal protein	
YNR037C	40S **ribosomal** protein S15e [SP:YN8L_YEAST]
YPL004C	YPL004C; Lpa13p
YNR071C	similar to UDP-glucose 4-epimerase GAL10P [SP:YN9A_YEAST]
YDL230W	PTP1; protein-tyrosine phosphatase 1 (ptpase 1) [EC:3.1.3.48] [SP:PTP1_YEAST]
YGL222C	unknown [SP:YGX2_YEAST]
YJR119C	unknown [SP:YJ89_YEAST]
YDR008C	unknown
YDR203W	unknown
YOR097C	unknown
YJR123W	RPS5; 40S **ribosomal** protein S5e [SP:RS5_YEAST]
YDL075W	RPL31A; 60S **ribosomal** protein L31e [SP:RL31_YEAST]
YHL033C	RPL8A; 60S **ribosomal** protein L7Ae [SP:RL4A_YEAST]
YDL208W	NHP2; high mobility group-like nuclear protein [SP:NHP2_YEAST]
YMR202W	ERG2; C-8 sterol isomerase [SP:ERG2_YEAST]
YNL146W	unknown [SP:YNO6_YEAST]
YBL072C	RPS8A; 40S **ribosomal** protein S8e [SP:RS8_YEAST]
YDL081C	RPP1A; 60S acidic **ribosomal** protein LP1 [SP:RLA1_YEAST]
YLR312C	unknown
YGL189C	RPS26A; 40S **ribosomal** protein S26e [SP:R26A_YEAST]
YGL030W	RPL30; 60S **ribosomal** protein L30e [SP:RL30_YEAST]
YDR529C	QCR7; ubiquinol-cytochrome c reductase subunit 7 [EC:1.10.2.2] [SP:UCR7_YEAST]
YOR167C	RPS28A; 40S **ribosomal** protein S28e [SP:RS28_YEAST]
YLR264W	RPS28B; 40S **ribosomal** protein S28e [SP:RS28_YEAST]
YLR340W	RPP0; 60S acidic **ribosomal** protein LP0
YDR382W	RPP2B; 60S acidic **ribosomal** protein LP2 [SP:RLA4_YEAST]
YGR085C	RPL11B; 60S **ribosomal** protein L11e [SP:RL11_YEAST]
YJR048W	CYC1; cytochrome c, iso-1 [SP:CYC1_YEAST]
(B) Translation	
YPR016C	CDC95; **translation** initiation factor 6 (eIF6)
YDR012W	RPL4B; 60S ribosomal protein L4e [SP:RL4B_YEAST]
YFL037W	TUB2; beta-tubulin [SP:TBB_YEAST]
YBR143C	SUP45, SUP1, SAL4; eukaryotic peptide chain release factor eRF subunit 1 [SP:ERF1_YEAST]
YNL062C	GCD10, TIF33; eukaryotic **translation** initiation factor eIF-3 gamma subunit [SP:IF33_YEAST]
YIL066C	RNR3; ribonucleoside-diphosphate reductase alpha chain [EC:1.17.4.1] [SP:RIR3_YEAST]
YGR204W	ADE3; methylenetetrahydrofolate dehydrogenase (NADP+) / methenyltetrahydrofolate cyclohydrolase / formate–tetrahydrofolate ligase [EC:1.5.1.5 3.5.4.9 6.3.4.3] [SP:C1TC_YEAST]
YPR041W	TIF5; eukaryotic **translation** initiation factor eIF-5 [SP:IF5_YEAST]
YDR224C	HTB1; histone H2B.1 [SP:H2B1_YEAST]
YGR222W	PET54; mitochondrial splicing protein and **translational** activator [SP:PT54_YEAST]
YAR042W	SWH1; probable NH-terminus of OSH1/SWH1 [SP:SWH1_YEAST]
YOL139C	CDC33, TIF45; eukaryotic **translation** initiation factor eIF-4E [SP:IF4E_YEAST]
YOR276W	CAF20, CAP20; mRNA CAP-binding protein (eIF4F), 20K subunit [SP:IF43_YEAST]
YPR163C	TIF3, STM1; eukaryotic **translation** initiation factor eIF-4B [SP:IF4B_YEAST]
YBR038W	CHS2; chitin synthase 2 [EC:2.4.1.16] [SP:CHS2_YEAST]
YKR026C	GCN3, AAS2, TIF221; **translation** initiation factor eIF-2B alpha subunit [SP:E2BA_YEAST]
YPR033C	HTS1; histidyl-tRNA synthetase [EC:6.1.1.21] [SP:SYH_YEAST]
(C) Transcription	
YOR344C	TYE7, SGC1; basic helix-loop-helix **transcription** factor [SP:TYE7_YEAST]
YGR151C	unknown [SP:YG3N_YEAST]
YNL300W	unknown [SP:YN40_YEAST]
YOR062C	unknown
YMR039C	SUB1, TSP1; **transcriptional** coactivator [SP:SUB1_YEAST]
YIL004C	BET1; protein transport protein [SP:BET1_YEAST]
YIL005W	protein disulfide-isomerase [EC:5.3.4.1] [SP:YIA5_YEAST]
YIR017C	MET28; **transcriptional** activator of sulfur amino acid metabolism [SP:MT28_YEAST]
YKR101W	SIR1; silencing regulatory protein [SP:SIR1_YEAST]
YIL025C	unknown [SP:YIC5_YEAST]
YGL112C	TAF60; **transcription** initiation factor [SP:T2D5_YEAST]
YGL122C	NAB2; nuclear polyadenylated RNA-binding protein [SP:NAB2_YEAST]
YKL109W	HAP4; **transcriptional** activator [SP:HAP4_YEAST]
YGR236C	unknown [SP:YG4Z_YEAST]
YDR259C	YAP6; **transcription** factor of a fungal-specific family of bzip proteins
YMR118C	succinate dehydrogenase (ubiquinone) cytochrome b subunit precursor [EC:1.3.5.1] [SP:YM07_YEAST]
YIL084C	SDS3; **transcriptional** regulator [SP:SDS3_YEAST]
YNL202W	SPS19; peroxisomal 2,4-dienoyl-CoA reductase [SP:SP19_YEAST]
YOR358W	HAP5; **transcriptional** activator [SP:HAP5_YEAST]
YDR277C	MTH1; repressor of hexose transport genes [SP:MTH1_YEAST]
YOR028C	CIN5; **transcriptional** activator [SP:CIN5_YEAST]
YGR167W	CLC1; clathrin light chain [SP:CLC1_YEAST]
YHR058C	MED6; RNA polymerase II **transcriptional** regulation mediator [SP:MED6_YEAST]
YOL152W	FRE7; unknown
YDR423C	CAD1; **transcriptional** activator [SP:CAD1_YEAST]
YJR019C	TES1; peroxisomal acyl-CoA thioesterase [SP:YJY9_YEAST]
YGL166W	CUP2, ACE1; **transcriptional** activator protein ACE1 [SP:ACE1_YEAST]
YIL132C	unknown [SP:YIN2_YEAST]
YJR094C	IME1; **transcription** factor involved in meiosis [SP:IME1_YEAST]
YOL042W	unknown
YML015C	TAF40; **transcription** initiation factor TFIID subunit [SP:T2D7_YEAST]
YPR120C	CLB5; cyclin, B-type [SP:CGS5_YEAST]

Table 2: Correlated gene clusters related to “ribosomal”, “translation”, and “transcription”.

(A) Autophagy	
YMR159C	**APG16**; similar to human Sin3 complex component SAP18, possible coiled-coil protein [SP:YM34_YEAST]
YJR025C	BNA1, HAD1; 3-hydroxyanthranilate 3,4-dioxygenase [EC:1.13.11.6] [SP:3HAO_YEAST]
(B) Spindle pole body function	
YKR037C	**SPC34**; spindle pole body protein [SP:YK17_YEAST]
YCR082W	unknown [SP:YCX2_YEAST]
YLR423C	unknown
YMR124W	unknown [SP:YM11_YEAST]
YIL144W	**TID3**; Dmc1p interacting protein [SP:YIO4_YEAST]
YOR089C	VPS21, YPT51; GTP-binding protein [SP:YP51_YEAST]
(C) Vesicular transport	
YBL050W	**SEC17**; vesicular-fusion protein [SP:SC17_YEAST]
YDR178W	SDH4; succinate dehydrogenase membrane anchor subunit [EC:1.3.5.1] [SP:SDH4_YEAST]
YMR197C	**VTI1**; vesicle transport V-snare protein VTI1 [SP:VTI1_YEAST]
YOR036W	**PEP12**; syntaxin (T-SNARE), vacuolar [SP:PE12_YEAST]
YDR468C	**TLG1**; tSNARE that affects a late Golgi compartment
YGL044C	RNA15; component of the cleavage and polyadenylation factor CF I involved in pre-mRNA 3'-end processing [SP:RN15_YEAST]
YBL102W	**SFT2**; SFT2 protein [SP:SFT2_YEAST]
YNL133C	unknown [SP:YNN3_YEAST]
YOR220W	unknown
YIL004C	**BET1**; protein transport protein [SP:BET1_YEAST]
YIL005W	protein disulfide-isomerase [EC:5.3.4.1] [SP:YIA5_YEAST]
YMR039C	SUB1, TSP1; transcriptional coactivator [SP:SUB1_YEAST]
YLR324W	unknown
YDR453C	YDR453C; probable thiol-specific antioxidant protein 2 [SP:TSA2_YEAST]
YNL044W	**YIP3**; unknown [SP:YIPC_YEAST]
YGR192C	TDH3, GPD3; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P3_YEAST]
YJL052W	TDH1, GPD1, SSS2; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P1_YEAST]
YDR313C	**PIB1**; phosphatidylinositol(3)-phosphate binding protein
YGR042W	unknown [SP:YG1T_YEAST]
YGL198W	unknown [SP:YGU8_YEAST]
YNL216W	RAP1, GRF1; DNA-binding protein with repressor and activator activity [SP:RAP1_YEAST]
YJL036W	**SNX4**; unknown [SP:YJD6_YEAST]
YDR473C	PRP3; essential splicing factor
YKR014C	**YPT52**; GTP-binding protein of the rab family [SP:YP52_YEAST]
YKL035W	UGP1; UTP–glucose-1-phosphate uridylyltransferase [EC:2.7.7.9] [SP:UDPG_YEAST]
YFL054C	unknown [SP:YFF4_YEAST]
YDR425W	unknown
YPL280W	unknown

Table 3: Correlated gene clusters related to the genes listed in Ito et al. [3].

extracted correlated gene clusters. In this section, we extend the hyperedges so that they can contain datasets from multiple organisms.

It is well known that in prokaryotic genomes, such as in *E.coli*, functionally related genes are often located continuously on the chromosome constituting an operon. Unfortunately, this is not usually the case for eukaryotic genomes including *S.cerevisiae*. However, if we can define functional identity of genes between the two species, the operon information in *E.coli* may be utilized for identifying functional links in *S.cerevisiae*.

Fig. 4 shows a schematic picture. To connect two organisms, we introduce a mapping from a set of genes in one organism to another based on the sequence similarities (orthologous relationships) as follows:

$$F_{SS} : S(G_1) \rightarrow S(G_2) \tag{2}$$

where $S(G_1)$ and $S(G_2)$ denote the sets of genes in the graphs G_1 and G_2 corresponding to two organisms, respectively. We use the criterion of bidirectional best hits to define orthologs when two genomes are compared by the Smith-Waterman algorithm at the amino acid sequence level with the threshold similarity score of 70. To characterize genes of an organism, its genes $S(G_1)$ are once mapped to the nodes of the graph G_2 that encodes functional orthologs in another organism. After that, we compare G_2 and an additional graph G_3 of the original organism instead of comparing G_1 and G_3 directly.

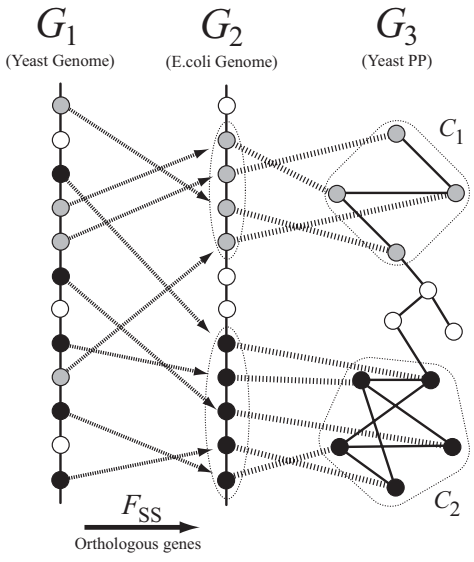

Figure 4: Mapping of genes.

Suppose that G_1 and G_3 are the binary relationships among *S.cerevisiae* genes with respect to

(A) Correlated gene clusters with single EC number

YJL140W	RPB4; DNA-directed RNA polymerase II 32 kD polypeptide [EC:2.7.7.6] [SP:RPB4_YEAST]
YKL144C	RPC25, YKL1, UNF1; DNA-directed RNA polymerase III 25 kD polypeptide [EC:2.7.7.6] [SP:RPCY_YEAST]
YOR116C	RPO31, RPC1, RPC160; DNA-directed RNA polymerase III largest subunit [EC:2.7.7.6] [SP:RPC1_YEAST]
YOR210W	RPB10; DNA-directed RNA polymerases I, II, and III 8.3 kD polypeptide [EC:2.7.7.6] [SP:RPBX_YEAST]
YJR063W	RPA12, RRN4; DNA-directed RNA polymerase I 13.7 kD polypeptide [EC:2.7.7.6] [SP:RPA9_YEAST]
YDR156W	RPA14; DNA-directed RNA polymerase I 14 kD polypeptide [EC:2.7.7.6] [SP:RPA8_YEAST]
YAR071W	PHO11; acid phosphatase [EC:3.1.3.2] [SP:PPAB_YEAST]
YHR215W	PHO12; acid phosphatase [EC:3.1.3.2] [SP:PPAC_YEAST]
YBR278W	DPB3; DNA polymerase epsilon, subunit C [EC:2.7.7.7] [SP:DPB3_YEAST]
YCR014C	POL4; DNA polymerase IV [EC:2.7.7.7] [SP:DPO4_YEAST]
YDL066W	IDP1; isocitrate dehydrogenase (NADP+), mitochondrial [EC:1.1.1.42] [SP:IDHP_YEAST]
YIR037W	HYR1; glutathione peroxidase [EC:1.11.1.9] [SP:GSHJ_YEAST]
YDL168W	SFA1; formaldehyde dehydrogenase (glutathione) / long-chain alcohol dehydrogenase [EC:1.2.1.1 1.1.1.1] [SP:FADH_YEAST]
YMR083W	ADH3; alcohol dehydrogenase [EC:1.1.1.1] [SP:ADH3_YEAST]
YDR226W	ADK1; adenylate kinase [EC:2.7.4.3] [SP:KAD1_YEAST]
YER170W	ADK2, PAK3; adenylate kinase [EC:2.7.4.3] [SP:KAD2_YEAST]
YDR256C	CTA1; catalase [EC:1.11.1.6] [SP:CATA_YEAST]
YGR088W	CTT1; catalase [EC:1.11.1.6] [SP:CATT_YEAST]
YGL070C	RPB9; DNA-directed RNA polymerase II 14.2 kD polypeptide [EC:2.7.7.6] [SP:RPB9_YEAST]
YOR224C	RPB8; DNA-directed RNA polymerase I, II, III 16 KD subunit [EC:2.7.7.6] [SP:RPB8_YEAST]
YGR192C	TDH3, GPD3; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P3_YEAST]
YJL052W	TDH1, GPD1, SSS2; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P1_YEAST]
YGR240C	PFK1; 6-phosphofructokinase [EC:2.7.1.11] [SP:K6P1_YEAST]
YMR205C	PFK2; 6-phosphofructokinase [EC:2.7.1.11] [SP:K6P2_YEAST]
YER081W	SER3; D-3-phosphoglycerate dehydrogenase [EC:1.1.1.95] [SP:SERX_YEAST]
YIL074C	SER33; 3-phosphoglycerate dehydrogenase [EC:1.1.1.95]
YLR157C	ASP3-2; L-asparaginase [EC:3.5.1.1] [SP:ASG2_YEAST]
YLR158C	ASP3-3; L-asparaginase [EC:3.5.1.1] [SP:ASG2_YEAST]
YOR224C	RPB8; DNA-directed RNA polymerase I, II, III 16 KD subunit [EC:2.7.7.6] [SP:RPB8_YEAST]
YOR340C	RPA43, RRN12; DNA-dependent RNA polymerase 36 kD polypeptide [EC:2.7.7.6] [SP:RPA4_YEAST]
YIL078W	THS1; threonyl-tRNA synthetase, cytoplasmic [EC:6.1.1.3] [SP:SYTC_YEAST]
YKL194C	MST1; threonyl-tRNA synthetase, mitochondrial [EC:6.1.1.3] [SP:SYTM_YEAST]

(B) Correlated gene clusters with multiple EC numbers

YDR321W	ASP1; L-asparaginase [EC:3.5.1.1] [SP:ASG1_YEAST]
YPR145W	ASN1; asparagine synthase (glutamine-hydrolysing) [EC:6.3.5.4] [SP:ASN1_YEAST]
YER086W	ILV1; threonine dehydratase [EC:4.2.1.16] [SP:THDH_YEAST]
YBR145W	ADH5; alcohol dehydrogenase [EC:1.1.1.1] [SP:ADH5_YEAST]
YER073W	ALD3; aldehyde dehydrogenase [EC:1.2.1.3] [SP:DHA5_YEAST]
YER043C	SAH1; adenosylhomocysteinase [EC:3.3.1.1] [SP:SAHH_YEAST]
YFR055W	cystathionine beta-lyase [EC:4.4.1.8] [SP:METC_YEAST]
YFL022C	FRS2; phenylalanyl-tRNA synthetase alpha chain [EC:6.1.1.20] [SP:SYFB_YEAST]
YKL106W	AAT1; aspartate aminotransferase [EC:2.6.1.1] [SP:AATM_YEAST]
YFR047C	nicotinate-nucleotide pyrophosphorylase (carboxylating) [EC:2.4.2.19] [SP:NADC_YEAST]
YLR209C	YLR209C; probable purine nucleoside phosphorylase [EC:2.4.2.1] [SP:PNPH_YEAST]
YIL066C	RNR3; ribonucleoside-diphosphate reductase alpha chain [EC:1.17.4.1] [SP:RIR3_YEAST]
YNR003C	RPC34; DNA-directed RNA polymerase III, 34 KD subunit [EC:2.7.7.6] [SP:RPC6_YEAST]
YIL066C	RNR3; ribonucleoside-diphosphate reductase alpha chain [EC:1.17.4.1] [SP:RIR3_YEAST]
YOR074C	CDC21; thymidylate synthase (TS) [EC:2.1.1.45] [SP:TYSY_YEAST]
YKL106W	AAT1; aspartate aminotransferase [EC:2.6.1.1] [SP:AATM_YEAST]
YOR374W	ALD4; aldehyde dehydrogenase [EC:1.2.1.3] [SP:DHA4_YEAST]
YDR148C	KGD2; 2-oxoglutarate dehydrogenase E2 component (dihydrolipoamide succinyltransferase) [EC:2.3.1.61] [SP:ODO2_YEAST]
YFL018C	LPD1, DHLP1; dihydrolipoamide dehydrogenase [EC:1.8.1.4] [SP:DLDH_YEAST]
YFR055W	cystathionine beta-lyase [EC:4.4.1.8] [SP:METC_YEAST]
YGR124W	ASN2; asparagine synthase (glutamine-hydrolysing) [EC:6.3.5.4] [SP:ASN2_YEAST]
YGL070C	RPB9; DNA-directed RNA polymerase II 14.2 kD polypeptide [EC:2.7.7.6] [SP:RPB9_YEAST]
YKL067W	YNK1; nucleoside-diphosphate kinase [EC:2.7.4.6] [SP:NDK_YEAST]
YHR216W	PUR5; IMP dehydrogenase [EC:1.1.1.205] [SP:IMH1_YEAST]
YLR209C	YLR209C; probable purine nucleoside phosphorylase [EC:2.4.2.1] [SP:PNPH_YEAST]
YBR035C	PDX3; pyridoxamine 5'-phosphate oxidase [EC:1.4.3.5] [SP:PDX3_YEAST]
YLR058C	SHM2; serine hydroxymethyltransferase, cytosolic (glycine hydroxymethyltransferase) [EC:2.1.2.1] [SP:GLYC_YEAST]
YGR204W	ADE3; methylenetetrahydrofolate dehydrogenase (NADP+) / methenyltetrahydrofolate cyclohydrolase / formate–tetrahydrofolate ligase [EC:1.5.1.5 3.5.4.9 6.3.4.3] [SP:C1TC_YEAST]
YOR074C	CDC21; thymidylate synthase (TS) [EC:2.1.1.45] [SP:TYSY_YEAST]

Table 4: Correlated gene clusters found in two-hybrid and pathway datasets.

sce:YBR020W	**GAL1; galactokinase [EC:2.7.1.6] [SP:GAL1_YEAST]**
eco:b0757	galK, galA; galactokinase [EC:2.7.1.6] [SP:GAL1_ECOLI]
sce:YBR018C	**GAL7; galactose-1-phosphate uridylyltransferase [EC:2.7.7.10] [SP:GAL7_YEAST]**
eco:b0758	galT, galB; galactose-1-phosphate uridylyltransferase [EC:2.7.7.10] [SP:GAL7_ECOLI]
sce:YCR021C	**HSP30; heat shock protein [SP:HS30_YEAST]**
eco:b2611	hypothetical protein
sce:YOR232W	**MGE1; GRPE protein homolog precursor [SP:GRPE_YEAST]**
eco:b2614	grpE; heat shock protein grpE (heat shock protein b25.3) (HSP24) [SP:GRPE_ECOLI]
sce:YMR058W	**FET3; iron transport multicopper oxidase precursor [EC:1.-.-.-] [SP:FET3_YEAST]**
eco:b0123	yacK; probable 53.4 kD blue-copper protein yacq precursor [SP:YACK_ECOLI]
sce:YNL036W	**NCE103; involved in non-classical protein export pathway [SP:NCE3_YEAST]**
eco:b0126	yadF; hypothetical 25.1 kD protein in hpt-panD intergenic region [SP:YADF_ECOLI]
sce:YDR226W	**ADK1; adenylate kinase [EC:2.7.4.3] [SP:KAD1_YEAST]**
eco:b0474	adk, plsA, dnaW; adenylate kinase [EC:2.7.4.3] [SP:KAD_ECOLI]
sce:YOR176W	**HEM15; ferrochelatase [EC:4.99.1.1] [SP:HEMZ_YEAST]**
eco:b0475	hemH, popA, visA; ferrochelatase (protoheme ferro-lyase) (hemE synthetase) [EC:4.99.1.1] [SP:HEMZ_ECOLI]
sce:YGR263C	**unknown [SP:YG5J_YEAST]**
eco:b0476	aes; acetyl esterase [EC:3.1.1.-] [SP:AES_ECOLI]
sce:YPR125W	**MRS7; suppressor of mrs2-1 mutation**
eco:b1384	feaR, maoR, maoB; transcriptional activator feaR [SP:FEAR_ECOLI]
sce:YOR374W	**ALD4; aldehyde dehydrogenase [EC:1.2.1.3] [SP:DHA4_YEAST]**
eco:b1385	feaB, padA, maoB; phenylacetaldehyde dehydrogenase (PAD) [EC:1.2.1.39] [SP:FEAB_ECOLI]
sce:YPL005W	**YPL005W; Lpa12p**
eco:b1696	putative ARAC-type regulatory protein
sce:YGR207C	**ETF-BETA; electron transfer flavoprotein beta-subunit [SP:ETFB_YEAST]**
eco:b1697	ydiQ; putative electron transfer flavoprotein subunit ydiq [SP:YDIQ_ECOLI]
sce:YPL252C	**YAH1; similar to adrenodoxin and ferrodoxin**
eco:b2525	fdx; ferredoxin, 2fe-2s [SP:FER_ECOLI]
sce:YGL018C	**JAC1; molecular chaperone [SP:YGB8_YEAST]**
eco:b2527	hscB; chaperone protein hscB (hsc20) [SP:HSCB_ECOLI]
sce:YMR253C	**unknown [SP:YM87_YEAST]**
eco:b3184	yhbE; hypothetical 35.0 kD protein in dacB-rpmA intergenic region (F321) [SP:YHBE_ECOLI]
sce:YNL005C	**MRPL2; mitochondrial ribosomal protein L2 precursor [SP:RM02_YEAST]**
eco:b3185	rpmA; 50S ribosomal protein L27 [SP:RL27_ECOLI]
sce:YBR034C	**HMT1, ODP1, RMT1; hnRNP arginine N-methyltransferase [EC:2.1.1.-] [SP:HMT1_YEAST]**
eco:b3259	prmA; ribosomal protein l11 methyltransferase [EC:2.1.1.-] [SP:PRMA_ECOLI]
sce:YLR401C	**unknown [SP:YL01_YEAST]**
eco:b3260	yhdG; hypothetical 35.9 kD protein in pmra-fis intergenic region (ORF1) [SP:YHDG_ECOLI]
sce:YDR268W	**MSW1; mitochondrial tryptophanyl-tRNA synthetase [EC:6.1.1.2] [SP:SYWM_YEAST]**
eco:b3384	trpS; tryptophanyl-tRNA synthetase [EC:6.1.1.2] [SP:SYW_ECOLI]
sce:YOR131C	**unknown**
eco:b3385	gph; phosphoglycolate phosphatase [EC:3.1.3.18] [SP:GPH_ECOLI]
sce:YGR218W	**CRM1; chromosome region maintenance protein [SP:CRM1_YEAST]**
eco:b3438	gntR; gluconate utilization system gnt-I transcriptional repressor [SP:GNTR_ECOLI]
sce:YMR315W	**unknown [SP:YM94_YEAST]**
eco:b3440	yhhX; hypothetical 38.8 kD protein in gntR-ggt intergenic region (F345) [SP:YHHX_ECOLI]

Table 5: Correlated gene clusters obtained by multiple organism comparison.

the positions in the *S.cerevisiae* genome sequence and the protein-protein interactions by two-hybrid analysis, and G_2 is the binary relationships among *E.coli* genes with respect to the positions in the *E.coli* genome sequence. To extract correlated gene clusters in the *S.cerevisiae* two-hybrid dataset G_3 by using the *E.coli* genome dataset G_2, *S.cerevisiae* genes in G_1 are mapped to *E.coli* genes in G_2 (dashed arrows). Then, the nodes in G_2 and those in G_3 are connected by hyperedges (dashed lines), and by clustering those hyperedges as explained in the previous sections we obtain correlated gene clusters like C_1 and C_2 that contain sets of genes reserving their mutual relationships in G_2 and G_3.

Actually, we connected the *S.cerevisiae* two-hybrid dataset and the *E.coli* dataset via 934 homologous relationships between two genes of these organisms.[8] By clustering the hyperedges we found eleven correlated gene clusters as shown in Table 5 (execution time was 14 seconds). Each *S.cerevisiae* gene is attached by its *E.coli* homologue. Here, genes that interact through at most one intermediate gene were included in the same cluster (i.e., for the ith graph, setting $p_i = 2$). The result includes, for example, correlated gene clusters related to the galactose metabolism (sce:YBR020W (GAL1) and sce:YBR018C (GAL7)) and heat shock proteins (sce:YCR021C (HSP30) and sce:YOR232W (MGE1)).

4 Discussions

One crucial point of the current method is whether the graphs being compared really can provide biological information to classify genes. Even if the dataset is based on biological facts, we must con-

[8]We used "best-best" entries of the KEGG/SSDB database.
See http://ssdb.genome.ad.jp/ for details.

sider its appropriateness for this purpose. For example, when we compared the *S.cerevisiae* genome dataset and the *S.cerevisiae* two-hybrid dataset, we actually found a total 106 "correlated gene clusters". However, considering the characteristics of the eukaryote genome sequences, it may not be easy to interpret the result. For example, YOL147C and YOL148C are next to each other on the *S.cerevisiae* genome sequence, and they are also connected through one intermediate gene YPR086W in the two-hybrid dataset. The annotations of these genes are as follows:

sce:YOL147C PEX11; peroxisomal membrane protein [SP:PEXB_YEAST]
sce:YOL148C SPT20, ADA5; transcription factor [SP:SP20_YEAST]
sce:YPR086W SUA7; transcription initiation factor IIB [SP:TF2B_YEAST]

But, different from prokaryote genomes that contain functional links as represented by operons, it is not clear whether the inclusion of the *S.cerevisiae* genome dataset can improve the confidence of screening the two-hybrid dataset. It is biologically true that the two genes above are next to each other in the *S.cerevisiae* genome sequence, but it cannot support that the interaction detected by two-hybrid analysis is biologically meaningful.

Currently, the output of the method is just a list of correlated gene clusters. Deriving sub-networks that indicate how genes are connected in a correlated gene cluster is a next subject of our analysis. They may be found by gathering the shortest paths between the genes in a correlated gene clusters. Those sub-networks may reveal intermediate members which are not contained in the list of genes in the correlated gene clusters, but which may still be of interest.

Finally, in this paper, we have focused on whether or not two genes are connected by means of binary relationships. Now, we can extend the framework so that it can cope with graphs whose edges have weights according to similarity scores, binding constants, and other quantitative relationships. Our algorithm works for this purpose too, but we must consider normalization of edge weights among different kinds of graphs (e.g., corresponding to a genome and a pathway) since comparison between their absolute values do not always make sense.

Acknowledgements

Part of this work done by A.N. and S.G. was supported by Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Information Science" from the Ministry of Education, Culture, Sports, Science and Technology of Japan. The work was also supported by the Research for the Future Program of the Japan Society for the Promotion of Science. The computing resource was provided by the Supercomputer Laboratory, Bioinformatics Center, Institute for Chemical Research, Kyoto University.

References

[1] DeRisi, J.L., Iyer, V.R., and Brown, P.O., Exploring the metabolic and genetic control of gene expression on a genome scale, *Science*, 278(5338):680–686, 1997.

[2] Erlandsen, H., Abola, E.E., and Stevens, R.C., Combining structural genomics and enzymology: completing the picture in metabolic pathways and enzyme active sites, *Curr. Opin. Struct. Biol.*, 10(6):719–730, 2000.

[3] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y., A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc. Natl. Acad. Sci. USA*, 98(8):4569–4574, 2001.

[4] Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C., SCOP: A structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, 247:536–540, 1995.

[5] Ogata, H., Fujibuchi, W., Goto, S., and Kanehisa, M., A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters, *Nucleic Acids Res.*, 28:4021–4028, 2000.