Extraction of Correlated Gene Clusters from Multiple Graph Structures: Application

Shuichi Kawashima shuichi@kuicr.kyoto-u.ac.jp
Akihiro Nakaya nakaya@kuicr.kyoto-u.ac.jp
Yoshinori Okuji okuji@kuicr.kyoto-u.ac.jp
Susumu Goto goto@kuicr.kyoto-u.ac.jp
Minoru Kanehisa kanehisa@kuicr.kyoto-u.ac.jp

[1] Institute for Chemical Research, Kyoto University Gokasho, Uji, Kyoto 611-0011, Japan

Keywords: correlated gene cluster, binary relation, graph isomorphism, clustering

1 Introduction

Many biological processes in the living organism can be expressed as a network of molecular interactions. Ogata *et al.* developed a method to extract similar subgraphs between two graphs (networks) [2]. They defined the subgraph similarity as a correlated cluster. This method, named SIMIC, is devised for the purpose of detecting correlated gene clusters without taking the topology of the network into consideration. Furthermore, Nakaya *et al.* developed a method to compare an arbitrary number of networks by extending the SIMIC algorithm [1]. This method can be applied to various networks that can be denoted by a set of binary relations, such as gene orders in the genome, molecular interactions, gene sequence similarities, protein structural similarities and coregulation of gene expressions. Especially, the data on large scale gene expression profiles by DNA chips and microarrays is likely to contain the information that can be uncovered by this method in the view of the post-genome analysis of connecting unknown genes to known genes. In this study, we report an actual application of the method to several kinds of real biological data sets.

2 Method and Results

We applied the method to two dataset. One consists of the three networks, which are the genome, the metabolic pathway and the structural similarity at the protein fold level in *Escherichia coli.* The

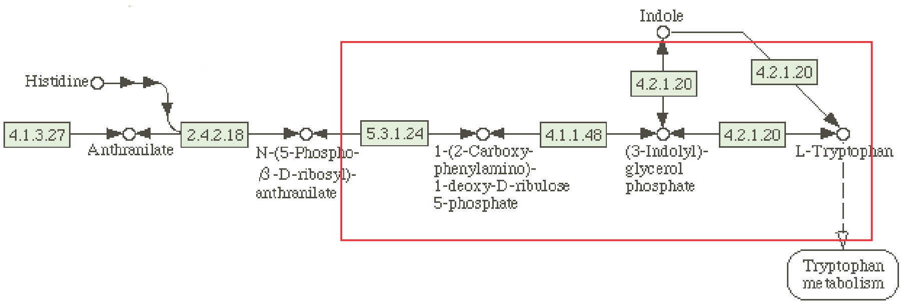
Indole
Histidine
4.2.1.20
4.2.1.20
4.1.3.27
Anthranilate
2.4.2.18
N-(5-Phospho-β-D-ribosyl)-anthranilate
5.3.1.24
1-(2-Carboxy-phenylamino)-1-deoxy-D-ribulose 5-phosphate
4.1.1.48
(3-Indolyl)-glycerol phosphate
4.2.1.20
L-Tryptophan
Tryptophan metabolism

Figure 1: Tryptophan biosynthesis pathway.

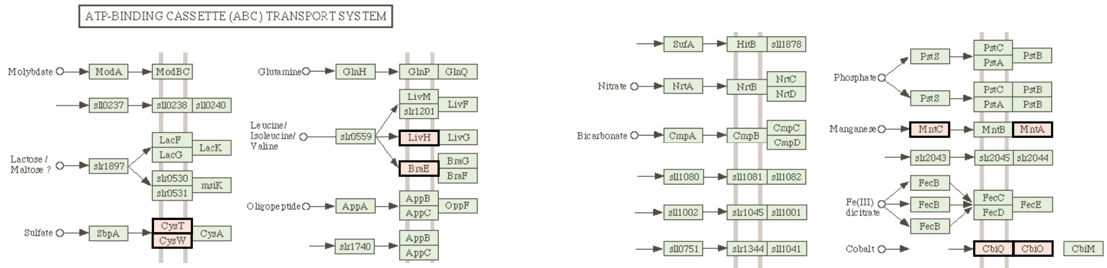

Figure 2: ABC transport system.

other consists of two networks, which are the metabolic/regulatory pathway and the similarity of gene expression profiles in *Synechocystis* sp. We obtained the data of the structure classification from the SCOP database and the genome and the pathway data from the KEGG database [3]. For gene expression profiles, we measured the similarity using the Euclid distance. Then, we performed the complete-linkage hierarchical clustering to the profiles using above distance as a metric. Therefore, each cluster composed a complete graph.

We detected a couple of correlated gene clusters from the three *E. coli* data sets. One of them is a part of tryptophan biosynthesis pathway as shown in Fig. 1. The cluster consists of three ORFs, b1260 (TrpA), b1261 (TrpB) and b1262 (TrpC). As is well known, these genes are located next to each other on the *E. coli* genome and composed the tryptophan operon. At the same time, these ORFs are classified as TIM-barrel structure proteins.

In the other example, we detected many correlated gene clusters from the two *Synechocystis* sp. data sets. Many among them are clusters involving metabolic pathways. It is difficult to judge whether they are biologically relevant. But, a few correlated clusters involving pathways of ABC transport systems seem to be of interest (Fig. 2).

3 Acknowledgments

This work was supported by grants from the Ministry of Education, Science, Sports and Culture, the Science and Technology Agency, and the Japan Society for the Promotion of Science. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

References

[1] Nakaya, A., Goto, S. and Kanehisa, M. *Genome Informatics*, 11, 2000.

[2] 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.

[3] Kanehisa M. and Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes *Nucleic Acids Res.*, 28:27–30, 2000.