Analysis of Binary Relations and Hierarchies of Enzymes in the Metabolic Pathways

Hiroyuki Ogata
ogata@kuicr.kyoto-u.ac.jp

Hidemasa Bono
bono@kuicr.kyoto-u.ac.jp

Wataru Fujibuchi
wataru@kuicr.kyoto-u.ac.jp

Susumu Goto
goto@kuicr.kyoto-u.ac.jp

and

Minoru Kanehisa
kanehisa@kuicr.kyoto-u.ac.jp

Institute for Chemical Research, Kyoto University
Gokasho, Uji, Kyoto 611 Japan

Abstract

In conjunction with a new database system that efficiently organizes the metabolic pathway data from various organisms, we are developing computational methodologies using binary relations and hierarchies of enzymes. Biological knowledge integrated in the system includes genes, gene products, chemical compounds, enzyme reactions and metabolic pathway diagrams. By automatically mapping the enzymes of a specific organism on the pathway diagrams, it becomes possible to visualize the characteristic features of the organism-specific metabolic pathways. With the aid of the computational methodology implemented in the system, it becomes again possible to analyze and investigate the pathways in terms of their function and evolution. In this paper, we describe the outline of the system and present new biological features of metabolic pathways revealed by the system.

1 Introduction

Representing the knowledge of the metabolic pathways on computational resources is an emergent and challenging subject in computational biology. First, the rapid progress in the experimental technology promotes the determination of genomic DNA sequences. After the first determination of the complete genome for a free-living organism,*Haemophilus influenzae*, in 1995[1], whole genomic sequences of several species have been successively determined[2, 3, 4, 5]. Now it is possible to systematically handle the whole gene products at once. Second, almost all the existing database in molecular biology have focused on the problems specific to individual genes or individual gene products. However, a cell function should be viewed as a systematic behavior of these molecules and their interactions. It is necessary to computerize the knowledge of the molecular interactions and the higher-level information of the system. Third, the metabolic pathways, a large part of which would be basic circuits for various cells from a wide range of organisms, have been studied fairly well. Thus, the development of the computational methodology for investigating the metabolic pathways should become also useful in tackling the other reaction pathways observed in cell cycle, signal transduction, *Drosophila* development, etc.

We have started a project named KEGG (Kyoto Encyclopedia of Genes and Genomes), to computerize the knowledge of the information pathways of biomolecules[6]. As an initial part of the project, we have collected and computerized the metabolic pathway data into the electronic form. The WWW implementation of KEGG serves several purposes. It allows researchers to examine the functional assignment of enzymes and gives a platform to predict the function of gene products. Visualization of possible metabolic pathways specific to each organism can be used for the interspecies comparison of the pathways. Since the data stored in KEGG has cross-links to the existing databases, the user would be quickly navigated to them by our DBGET integrated database system[7]. KEGG is tightly coupled with the LIGAND database[8], which consists of catalytic reaction formulas and structures of chemical compounds. It is therefore possible to calculate or deduce unknown pathways that is not explicitly drawn in the pathway diagrams stored in the system. It is also possible to systematically analyze the pathways themselves and the enzymes playing roles in them in terms of their function and evolution.

In the present study, we focus our analysis on the pairs of enzymes that appear on the metabolic pathways and the hierarchical classifications of enzymes. The enzyme-enzyme relationship is represented by, what is called, the binary relation. In order to characterize the organization of the metabolic pathways, we employ the enzyme-enzyme binary relations for calculating the pathway between two enzymes viewing the pathway diagrams as undirected graphs. The hierarchical classifications are considered as extensions of the binary relations, which are used for the prediction of enzyme functions.

2 Data and Methods

2.1 The WWW implementation of KEGG

At present the WWW implementation of KEGG consists of three types of data: pathways, genes and molecules.

2.1.1 Pathways

We collected the metabolic pathway data available in the published sources, mostly from Boehringer's biochemical pathway[9] and the compilation by the Japanese Biochemical Society[10], and reorganized them into about 80 sections of clickable graphic diagrams (Figure 1). The computerized diagram is not intended to capture the consensus pathway or an organism-specific pathway, but it represents a collection of all chemically feasible reaction pathways. Thus, the actual pathway in a cell or of a certain organism should correspond to a subset or a part of the diagram.

2.1.2 Genes

Thus far gene catalogs of seventeen species have been complied in KEGG. Each catalog consists of gene names, gene product names, links to other existing databases, and, if they are available, EC (Enzyme Commission) numbers for enzymes. The enzymes in the gene catalog for each organism are classified into categories, according to the classification of metabolic pathways that consists of 10 sections and about 80 subsections, using the EC number as a key.

2.1.3 Molecules

In the molecules section of the WWW implementation of KEGG, enzymes are hierarchically classified according to their functions and/or structures. The classification of EC numbers, which is based on the nature of chemical reactions catalyzed by enzymes, has four levels of hierarchy. The top level of hierarchy has six classes, i.e. oxidoreductase, transferase, hydrolase, lyase, isomerase and ligase. The deepest, fourth level corresponds to the list of roughly 3300 known catalytic reactions. The classification by sequence motifs in PROSITE[11] also reflects the enzyme function. The superfamily and the 3-D fold classifications represent the similarity based on the primary and the tertiary structures, respectively.

2.2 Binary relations and hierarchies in the path computation

We have proposed a basic strategy to represent and compute various types of data by binary relations[12]. This is suitable for computing pathways in KEGG and connecting related entries in DBGET.

We organize two types of binary relations for the pathway computation. One is the substrate-product relation in the form of

$$\text{cpd:C00118} \Rightarrow \text{cpd:C00111}$$

which is extracted from the LIGAND database. The other is the enzyme-enzyme relation,

$$\text{EC:4.1.3.18} \Rightarrow \text{EC:1.1.1.86}$$

which corresponds to a pair of enzymes connected in the pathway diagrams, i.e., the nearest neighbor enzymes. These binary relations are used for the calculation of pathways either from a compound to a compound or from an enzyme to an enzyme. Direct application of Dijkstra's algorithm or Floyd's algorithm enables us to compute the shortest path between two components in the metabolic pathway diagrams. In this paper, we used the shortest path length L derived from the enzyme-enzyme relations as a definition for the distance of two enzymes in the pathway diagrams. By definition the shortest path does not necessarily correspond to the successive flow of catalytic reactions that actually happens in living organisms. We used the definition only to capture the overall architecture of metabolic pathways.

Since the classifications described above reflect the similarity of enzymes that may or may not be detected by the sequence homology search, their use in the prediction of gene function and in the correction of functional assignments is one of the major challenges of the KEGG project. The classification of enzymes is a hierarchy, which may also be viewed as a collection of binary relations like the following form.

$$\text{motif:PS00011} \Leftrightarrow \text{EC:3.4.21.5}$$

This implies that there are multiple enzymes (EC numbers) that share the same motif. If we perform a join in relational operations using 'motif' as a key, we can deduce a new binary relation like the following form.

$$\text{EC:3.4.21.5} \Leftrightarrow \text{EC:3.4.21.21}$$

By using this kind of binary relation that represents a similarity of these two enzymes, we can search alternative enzymes in the pathway.

3 Prediction of enzyme function

3.1 Interspecies comparison of the pathways

The metabolic pathways can be different from organisms to organisms. By mapping the enzyme genes identified by the genome sequencing project on the pathway diagrams, an organism-specific pathway can be viewed as a sequence of marked boxes corresponding to the identified enzymes of a certain organism. Figure 1 shows the methionine biosynthesis pathway for *H. influenzae*.

This visualization technique enabled us to compare the metabolic pathways of different organisms. We compared the amino acid biosynthesis pathways between *Haemophilus influenzae* and *Escherichia coli* (Table 1). For three out of the twenty amino acids, the synthetic pathways were not found in either of the organisms. For two amino acids, the synthetic pathways were found only for *E. coli*. Although *E. coli* genome sequencing has not yet been completed, the latter case may need further investigation because the functional assignments of *H. influenzae* ORFs are mainly based on the homology search to other organisms, especially *E. coli*.

In *H. influenzae*, whose genomic sequence is completely determined, this visualization technique immediately identifies missing enzymes in the formation of a pathway. In Table 1, such missing enzymes are listed. It is interesting that transaminases (EC 2.6.1.-) are frequently found, which suggests a problem in the putative assignment of gene functions.

3.2 Searching alternative assignments for missing enzymes

Figure 1 shows the methionine metabolism pathway in which those enzymes identified in *H. influenzae* are marked. We demonstrate here an approach to analyze a putative missing enzyme taking cystathionine gamma-lyase (EC 4.4.1.1) in the pathway as an example.

The approach involves searching for alternative assignments using hierarchical classifications. We applied our system to finding the enzymes of *H. influenzae* that are similar to the missing enzyme (EC 4.4.1.1) by deducing from the binary relations representing the enzyme similarities according to the four classifications: EC numbers, superfamilies, motifs and folds. Two enzymes shown in Table 2 were obtained as possible alternatives. In this case, the EC number classification was used at the third level of numbering hierarchy.

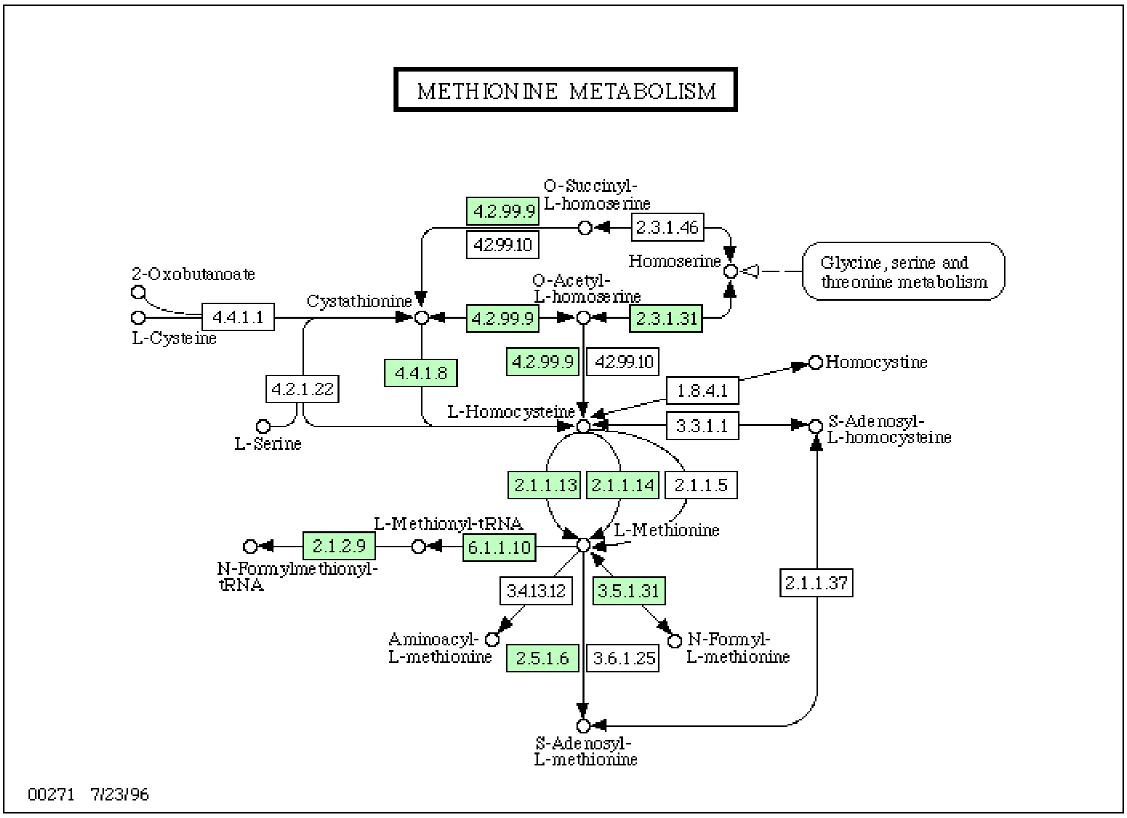

Figure 1: Methionine biosynthesis pathway. Those enzymes identified in *H. influenzae* are shaded.

Table 1: Comparison of amino acid biosynthesis pathways between *E. coli* and *H. influenzae*.

Amino acid	*H. influenzae*	*E. coli*	Missing in *H. influenzae*
Ala	x	x	2.6.1.2 or 2.6.1.12 or 2.6.1.18
Lys	x	x	1.4.1.16 or 2.6.1.17 or 3.5.1.47
Tyr	x	x	(2.6.1.57 and 1.3.1.43) or 1.4.3.2 or 2.6.1.5
Phe	x	o	1.4.3.2 or 1.4.1.20 or 2.6.1.5 or 2.6.1.57
Pro	x	o	2.6.1.13 or 4.3.1.12
Others*	o	o	

*Others: Asn, Asp, Arg, Cys, Gln, Glu, Gly, His, Ile, Leu, Met, Ser, Thr, Trp, Val

Table 2: The result of the alternatives search for a putative missing enzyme, the cystathionine gamma-lyase (EC 4.4.1.1), in the methionine biosynthesis pathway of *H. influenzae*

Alternative	Classification used
4.4.1.8	motif (Cys/Met metabolism enzymes pyridoxal-phosphate attachment site) & EC number classification
4.2.99.9	motif (Cys/Met metabolism enzymes pyridoxal-phosphate attachment site)

In general there are several possibilities that can cause missing enzymes. First, the organism actually does not have the enzyme. In the case of the methionine biosynthesis, methionine can be synthesized from homoserine as shown in figure 1 instead of the path containing the missing enzyme EC 4.4.1.1. Thus the lack of this enzyme may not be lethal for the organism. Second, the assignment of the function, namely the EC number, for the predicted coding region is wrong. In this case the alternatives as in Table 2 would be suggestive for the reassignment process. Third, the prediction of coding regions missed the functional open reading frame for the enzyme. However, this case might be rejected for cystathionine gamma-lyase, because TFASTA search[13] of the yeast protein sequence for this enzyme against the whole genomic DNA sequence of *H. influenzae* did not find such an ORF.

4 Heterogeneous organization of the pathways

4.1 Functional similarity and relative position of two enzymes

More than one thousand enzymes compose the metabolic pathways stored in the KEGG pathway diagrams. Taking two enzymes as constituents, we investigated and characterized the metabolic pathways by using again the hierarchy of enzymes.

We define a parameter, S, to represent the functional similarity of two enzymes according to the EC number classification. When two enzymes belong to the same EC number class up to the level k, where k is 1 to 3, the similarity class is defined by $S = k$. When they do not belong to the same class at the top level, $S = 0$ is assigned.

In order to capture the relative position of two enzymes in the pathway diagrams, we employed the shortest path length, L, calculated by Floyd's algorithm using the enzyme-enzyme binary relations as described in Data and Methods section. Among all the 1,031,766 pairs composed by 1,437 enzymes in the pathway diagrams, more than 90 % (954,451 pairs) could be connected up to certain lengths by the enzyme-enzyme binary relations extracted from the pathway diagrams. Figure 2 shows the frequency histogram. It is interesting to focus on the relative frequency of the similarity class S of the enzyme pairs in each pathway length L (Figure 3).

There is a strong tendency that the ratio of class $S = 2$ and 3 increases in accordance with the decreasing path length L. Chi-square statistical test indicates the dependency between the functional similarity class S and the path length L at a very high level of the significance ($P \ll 0.005$). Although this heterogeneity of the organization of metabolic pathways might come from various kinds of factors and would need further investigation, the fact that similar reactions (enzymes) tend to appear closely on the pathway immediately suggests the evolutional changes in the formation of metabolic pathways, such as gene duplication events.

4.2 Evolutionarily related enzymes

In order to estimate how much the correlation shown above is dependent on the evolutionary relationship of the enzymes, we used binary relations of enzymes derived by the superfamily classification.

Among the 954,451 pairs of enzymes, 114 pairs were extracted as belonging to the same superfamilies, i.e., having similar sequences. The frequency of these pairs against each path

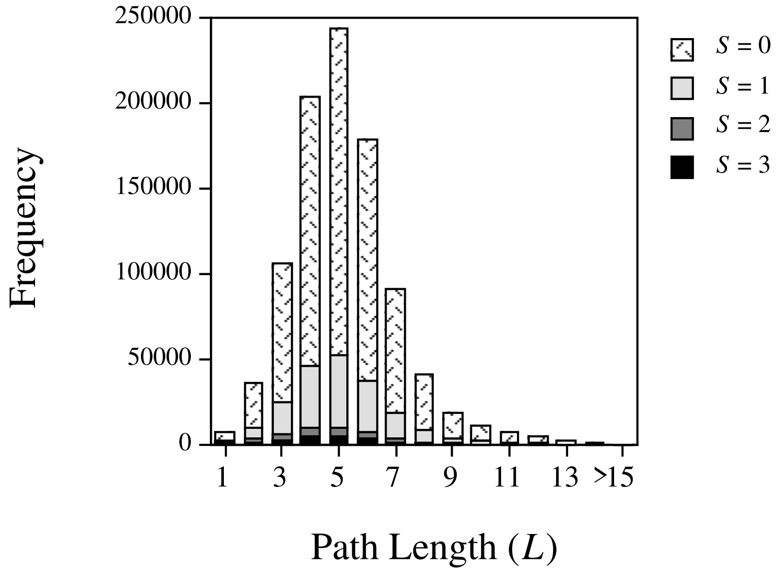

Figure 2: The distribution of path length L between arbitrary two enzymes. S represent the degree of functional similarity of two enzymes.

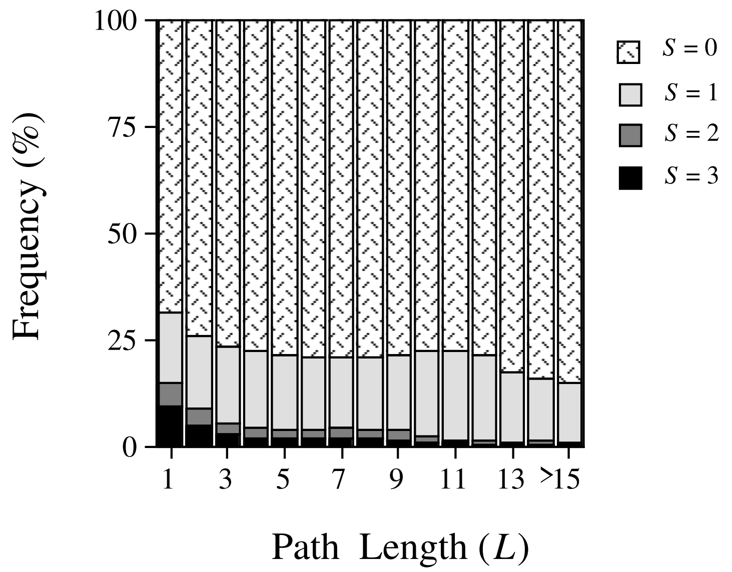

Figure 3: Relative frequency of the similarity class S in each path length L.

length L is plotted in Figure 4. It is interesting that many of these pairs are observed in much shorter path length positions in the metabolic pathway comparing to the distribution in Figure 2, even though the sample is not quite large. In addition to this tendency, the abundance of close functional similarities of enzyme pairs (S=2 and 3) is apparent, which is of course natural because they belong to the same superfamilies. Therefore, the correlation between functional similarity and the path length observed in Figure 3 seems to be the result of evolutional events. Further investigation will lead us to more detailed understanding of the organization of metabolic pathways.

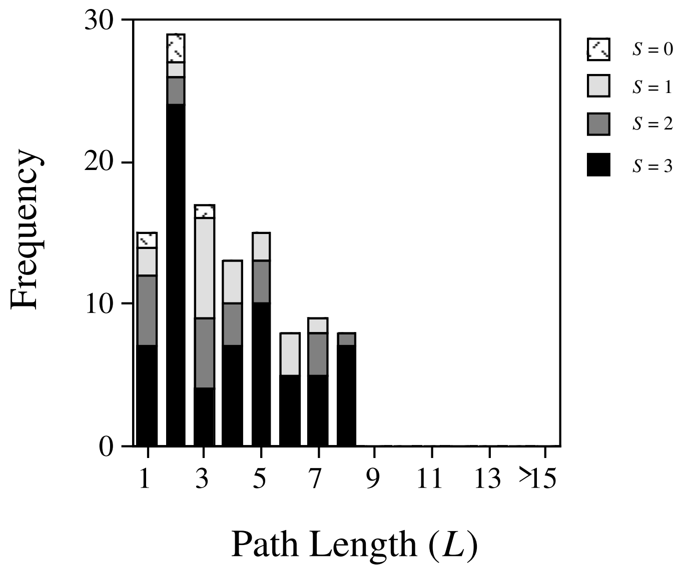

Figure 4: The distribution of the path length L between two evolutionally related enzymes.

5 Summary and Perspective

The knowledge of metabolic pathways is efficiently and usefully organized in KEGG for deduction from binary relations and hierarchies. The visualization and computation tools provided by KEGG are helpful for predicting and checking functional assignments of newly determined genes. In this paper, we have demonstrated how basic manipulations of binary relations and hierarchies of enzymes are used in the pathway analysis. Since the systematic analysis of reaction pathways is a new subject in computational biology, the KEGG metabolic pathway database will become a useful resource towards the progress of this research area.

Acknowledgement

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, 'Genome Science', from the Ministry of Education, Science, Sports and Culture of Japan.

The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

References

[1] Fleischmann, R.D., et al., "Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd," *Science*, Vol. 269, pp. 496-512, 1995.

[2] Fraser, C.M., et al., "The minimal gene complement of *Mycoplasma genitalium*," *Science*, Vol. 270, pp. 397-403, 1995.

[3] Kaneko, T., et al., "Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis*, sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions," *DNA Research*, Vol. 3, pp. 109-136, 1996.

[4] Bult, C.J., et al., "Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*," *Science*, Vol. 273, pp. 1058-1073, 1996.

[5] http://genome-www.stanford.edu/Saccharomyces/

[6] http://www.genome.ad.jp/kegg/kegg.html

[7] http://www.genome.ad.jp/dbget/dbget.links.html

[8] Suyama, M., Ogiwara, A., Nishioka, T., Oda, J., "Searching for amino acid sequence motifs among enzymes: the enzyme-reaction database," *Comput. Appli. Biosci.*, 9, 9-15 (1993)

[9] Gerhard, M. ed., Biological Pathways, Third Edition. Boehringer Mannheim (1992)

[10] Nishizuka, T. ed, Metabolic Maps. Biochemical Society of Japan (1980) (in Japanese)

[11] Bairoch, A., "PROSITE: a dictionary of sites and patterns in proteins," *Nucleic Acids Res.*, 19 Suppl, 2241-2245 (1991)

[12] Goto, S., Bono, H., Ogata, H., Fujibuchi, W., Nishioka, T., Sato, K., Kanehisa, M., "Organizing and computing metabolic pathway data in terms of binary relations," *Pacific Symposium on Biocomputing '97*, (1997) *in press*

[13] Pearson, W.R., Lipman, D.J., "Improved tools for biological sequence," *Proc. Natl. Acad. Sci. USA*, 85, 2444-2448 (1988)