IPSJ Transactions on Bioinformatics Vol. 2 15–24 (Mar. 2009)

Original Paper

Reaction Similarities Focusing Substructure Changes of Chemical Compounds and Metabolic Pathway Alignments

Yukako Tohsato[†1] and Yu Nishimura[†2]

Comparative analyses of enzymatic reactions provide important information on both evolution and potential pharmacological targets. Previously, we focused on the structural formulae of compounds, and proposed a method to calculate enzymatic similarities based on these formulae. However, with the proposed method it is difficult to measure the reaction similarity when the formulae of the compounds constituting each reaction are completely different. The present study was performed to extract substructures that change within chemical compounds using the RPAIR data in KEGG. Two approaches were applied to measure the similarity between the extracted substructures: a fingerprint-based approach using the MACCS key and the Tanimoto/Jaccard coefficients; and the Topological Fragment Spectra-based approach that does not require any predefined list of substructures. Whether the similarity measures can detect similarity between enzymatic reactions was evaluated. Using one of the similarity measures, metabolic pathways in *Escherichia coli* were aligned to confirm the effectiveness of the method.

1. Introduction

Comparative analysis of the enzymatic reactions provides essential information on both the evolution of organisms and on potential pharmacological targets[1], and there has been a great deal of research in this area in recent years. Enzymatic similarity is determined essentially based on sequence similarity between enzymes. However, it has been reported that comparisons based on sequence similarity are not always appropriate, because reaction similarities are not necessarily correlated with sequence similarity due to enzyme recruitment[2]. Therefore, emphasis is placed on comparison and investigation analysis results from variety of standpoints.

One approach involves protein domain families. Enzymes may be regarded as similar if they are evolutionarily related and this is the measure used by evolutionary trees, such as those in the Pfam protein domain family database. Alternatively, any enzymes that share structural elements may be regarded as similar; this is the basis of the CATH (Class, Architecture, Topology, Homologous superfamily) protein structure classification system. On the other hand, other classification systems consider enzyme function such as the Enzyme Commission (EC) system[3], which is based on the overall reaction catalyzed by the enzymes.

Although the EC classification system has proven its worth as a system for cataloguing and comparing the overall reactions catalyzed by enzymes, the assignment of the EC numbers is performed manually based on published experimental data on individual enzymes by the Joint Commission on Biochemical Nomenclature (JCBN) of the International Union of Pure and Applied Chemistry (IUPAC). The requirement of published articles on individual enzymes leaves many reactions unassigned, such as reactions known to be present in pathways and those inferred from chemical compounds. Therefore, there is a need for a similarity measure for enzymatic reactions[4].

Therefore, our previous study focused on the structural formulae of compounds ("compound structures")[5]. As identification of common isomorphic subgraphs between two compound structures is a NP-hard problem[6],[7], a number of methods for measuring the similarity between compound structures have been proposed. The main approach is the fingerprint-based comparison, which considers a molecule as a bit-string where each bit shows the presence or absence of either an atom or an important predefined molecular substructure called the key descriptor or finger[8]. We defined the enzymatic similarity between two substrate compounds and two product compounds based on those bit-strings. However, with the proposed method it is difficult to measure the reaction similarity, when the formulae of the compounds that constitute each reaction are completely different[5].

In this study, modified structural components ("component structures") were extracted from compound structures which constitute enzymatic reactions using RPAIR data[7],[9] in the Kyoto Encyclopedia of Genes and Genomes database

†1 Department of Bioscience and Bioinformatics, College of Life Sciences, Ritsumeikan University
†2 Information Science and Systems Engineering, Graduate School of Science and Engineering, Ritsumeikan University

(KEGG)[10]. Two reaction similarity scores were defined based on the extracted component structures, and these scoring systems were evaluated in comparison to EC classification hierarchy. The whole metabolic pathways were aligned using one of the reaction similarity scoring systems to find "pathway duplication". Pathway duplication suggested that evolution has occurred through duplication of the genes encoding proteins within a pathway[11,12]. This report presents the results obtained by applying the proposed method to actual metabolic pathway data in *E. coli* K-12 MG1655.

2. Materials and Methods

2.1 Extraction of Substructure Changes in Chemical Compounds

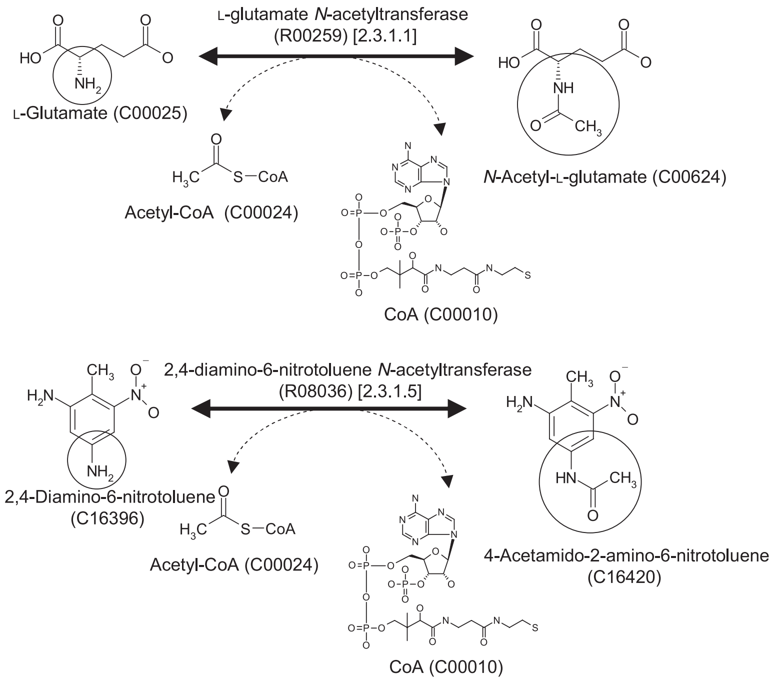
RPAIR stores chemical transformation patterns of structures that may occur between two compounds in a single reaction, which aligns pairs of compounds that have atoms or atom groups in common on two sides of a reaction[7,9]. The reactions are linked to its related entries in KEGG[10]. Using RPAIR, the atoms or atom groups can be divided into two groups, with and without different structures, depending on the structure alignment. Substructures that change within chemical compounds were extracted based on the RPAIR, and the extracted component data were reconstructed in MOL file format. An extracted substructure of a chemical compound is called a "component".

For example, in **Fig. 1**, the pairs of compounds (C00025, C00624) and (C16396, C16420) are major flows in the reactions R00259 and R08036, respectively, based on KEGG. Atom groups surrounded by circles indicate the modified structure components by each enzyme.

The KEGG database integrates the Reaction Classification (RC) system, which categorizes chemical transformation patterns in enzymatic reactions based on RPAIR[9]. However, in the RC system, comparison of the modified structure components is not considered.

2.2 Component Similarity and Reaction Similarity

Two methods were used in this study, *i.e.*, the fingerprint-based method[8] and the topological fragment spectra method[13,14], to calculate similarity between two extracted substructure changes of chemical compounds. The methods were extended to overall similarity measures for two enzymatic reactions.



L-glutamate *N*-acetyltransferase
(R00259) [2.3.1.1]
L-Glutamate (C00025)
Acetyl-CoA (C00024)
CoA (C00010)
N-Acetyl-L-glutamate (C00624)
2,4-diamino-6-nitrotoluene *N*-acetyltransferase
(R08036) [2.3.1.5]
2,4-Diamino-6-nitrotoluene
(C16396)
Acetyl-CoA (C00024)
CoA (C00010)
4-Acetamido-2-amino-6-nitrotoluene
(C16420)

Fig. 1 Enzymatic reactions of KEGG entries R00256, L-glutamate *N*-acetyltransferase, and R08036, 2,4-diamino-6-nitrotoluene *N*-acetyltransferase. The enzyme names are shown above their KEGG reaction IDs and EC numbers (EC numbers are given in brackets). The structural components modified by the enzymes are circled between their substrates and products (KEGG compound IDs are given in parentheses). Bold arrows are main reactions, and arrows indicated broken arrows are sub-reactions based on KEGG.

2.2.1 Fingerprint-based Approach

MACCS key was used[15], which is a fingerprint proposed by MDL and is one of the most widely used tools for computational screening of compound databases. This structure representation based on compound fragments is constructed by a string of keysets, which indicates whether a fragment of a specific substructure exists in the compound. Fragments of chemical structures can be coded in binary keys, which are presented as sequences of 0s and 1s (bit-strings). Here, 0 represents a fragment that does not exist in the structure; otherwise, the bit is 1, which indicates that the fragment exists. Specifically, an encoded bit-string has a length of 166 bits. The MACCS 166 key is designed for database indexing to increase speed in substructure searches, as described previously[15]. The key is not sensitive to 3D conformations, such as geometrical isomers; although maleic acid and fumaric acid are *cis-trans* isomers, the bit-strings are the same.

To measure the similarity between two structural formulae using the described fingerprint, a number of similarity measures have been proposed[8]. A widely used similarity measure, called the Tanimoto (Jaccard) coefficient, was employed:

$$F(c_1, c_2) = c/(a + b - c) \tag{1}$$

where a is the number of 1s of the bit-string of component c_1, b is the number of 1s of that of component c_2, and c is the number of 1s common to both c_1 and c_2. By definition, $0 \leq F(c_1, c_2) \leq 1$; the closer to 1, the higher the degree of similarity between the two bit-strings is, while the closer to 0, the lower the degree of similarity between the two bit-strings is.

The reaction similarity $PF(r_1, r_2)$ between the reaction (c_{11}, c_{12}) and the reaction (c_{21}, c_{22}) is calculated by the average of the component similarities of corresponding compounds as follows:

$$PF(r_1, r_2) = (F(c_{11}, c_{21}) + F(c_{12}, c_{22}))/2 \tag{2}$$

where the pairs of compounds (c_{11}, c_{12}) and (c_{21}, c_{22}) are major flows in the reactions r_1 and r_2, respectively. For example, changes in the compound structure catalyzed by L-glutamate N-acetyltransferase and 2,4-diamino-6-nitrotoluene N-acetyltransferase are identical $PF = 1$ (Fig. 1). In the case of a reaction with a number of main flows, the combination is selected that yields the highest score.

2.2.2 Topological Fragment Spectra-based Approach

The Topological Fragment Spectra (TFS)-based approach allows description of the topological structure profile of a molecule and does not require any predefined list of substructures. This approach is based on enumeration and numerical characterization of all possible substructures from a chemical structure[13),14)], and involves two main steps: (1) enumeration of all possible substructures in each chemical structure, and (2) numerical characterization of the substructures (**Fig. 2**). Figure 2 shows an example of enumeration of all possible substructures of 2-methylbutane and generation of vector data c = (3, 1, 2, 2, 2, 3, 3). Here, all of the substructures are characterized by the sum of degrees of the nodes composing each subgraph. Generally, the computational time required for the exhaustive enumeration of all possible substructures from a chemical structure is often very large, especially for molecules that contain highly fused rings. To avoid these problems, this method employed a subspectral approach, in which every TFS is described with structural fragments of the specified size or less. Takahashi, et al. proposed several types of TFS that can be defined with different graph representation schemes[13),14)]. The present study used the simplest approach in which the multiplicities of the chemical bounds and atomic types are ignored. All fragments with a size of 5 or less were used for generating TFS. The extracted components were converted to vector data.

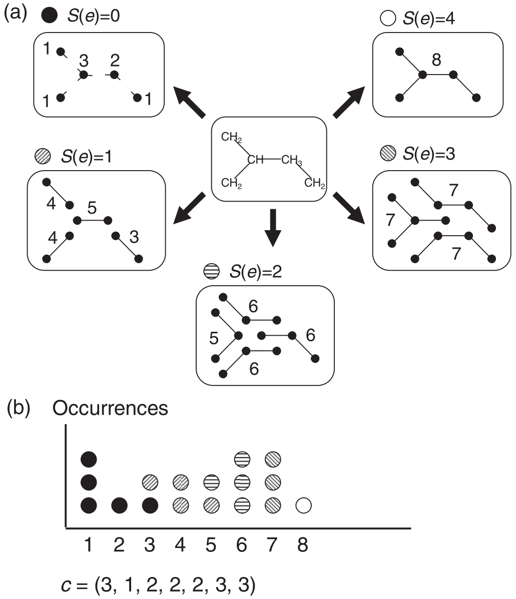


Fig. 2 Example of the procedure of TFS generation. (a) All possible substructures from a chemical structure are enumerated. $S(e)$ is the size (number of edges) of substructure to be enumerated. Every substructure is characterized by the sum of the nodes composing each subgraph. (b) The results are represented by vector data.

The degrees of dissimilarity between the vector data $(c_{11}, c_{12}, \cdots, c_{1m})$ of component c_1 and the vector data $(c_{21}, c_{22}, \cdots, c_{2m})$ of component c_2 is calculated using the Euclidean distance as follows:

$$T(c_1, c_2) = \sqrt{\sum (c_{1k} - c_{2k})^2} \tag{3}$$

The degree of dissimilarity using the Euclidean distance tends to become larger for pairs of vector data that are less similar.

The reaction dissimilarity $PT(r_1, r_2)$ between the reactions (c_{11}, c_{12}) and (c_{21}, c_{22}) is calculated by the average of the component dissimilarities of corresponding compounds as follows:

$$PT(r_1, r_2) = (T(c_{11}, c_{21}) + T(c_{12}, c_{22}))/2 \tag{4}$$

For example, since changes in the compound structure catalyzed by L-glutamate *N*-acetyltransferase and 2,4-diamino-6-nitrotoluene *N*-acetyltransferase are identical, $PT = 0$ (Fig. 1).

2.3 Pathway Alignment and Statistical Significance of Alignments

The local alignment algorithm based on dynamic programming (Smith and Waterman algorithm[16]) was extended for alignment between two metabolic pathways. In this study, a metabolic pathway from metabolite c_1 to c_m was defined as a sequence $(c_1, c_2)(c_2, c_3) \cdots (c_{m-1}, c_m)$ of pairs of compounds. It was identified with a sequence $r_1, r_2, \cdots, r_m$ of biochemical reactions adjacent to each other. The length of the pathway is the number of pairs of compounds.

For reaction sequences $r_{11}, r_{12}, \cdots, r_{1m}$ and $r_{21}, r_{22}, \cdots, r_{2n}$ as input, the alignment algorithm uses a matrix M. Let $M(i, j)$ $(0 \leq i \leq m, 0 \leq j \leq n)$ be a matrix initially filled with zeroes. The local alignment based on dynamic programming arranges the elements in each pair of sequences in two dimensions, and fills the matrix from left to right and top to bottom based on the following recursive relation transform M:

$$M[i, j] = \max \begin{cases} M[i-1, j-1] + S(r_i, r_j) \\ M[i-1, j] - d, \\ M[i, j-1] - d, \\ 0, \end{cases} \tag{5}$$

Thus, the alignment algorithm was extended by viewing $S(r_i, r_j)$ as reaction similarity. When a diagonal arrow is selected, the similarity score between two reactions corresponding to the arrow is added. When a left-to-right or top-to-bottom arrow is selected, a gap penalty is added. This algorithm is the same as that described previously[5] except for the definition of reaction similarity.

Here, the mean S_{mean} and standard deviation S_σ of all pair of reactions are calculated, and reaction similarities S_r are corrected as follows:

$$S = (S_r - S_{mean})/S_\sigma \tag{6}$$

with the gap penalty set to –0.001.

In addition, local alignment scores are generally dependent on the length. The Z-value of an alignment is computed using sample data. The Z-value is defined as:

$$Z = (X_{align} - X_{mean})/X_\sigma \tag{7}$$

where X_{align} is the alignment score. X_{mean} and X_σ are the mean and standard deviation, respectively, of a large number of random sequences of the same length. Finally, each individual Z-value is converted into a P-value, defined as the probability of obtaining a value of the test statistic that is at least as extreme as that obtained for the sample data, and the statistical significance of an alignment is evaluated.

3. Results and Discussions

3.1 Comparison among Reaction Similarity Measures

A total of 19,805 substructures that change within chemical compounds based on the RPAIR from the KEGG database ver. 46.0 were extracted, and the extracted components were reconstructed in MOL file format. The files were converted to SMILES strings using Open Babel[17]. Bit-string data were generated from the SMILES strings using the Fingerprint Module of MESA[18], which generated 164 bit-strings from SMILES strings input. The bit-strings are a public subset of 166 MACCS keys.

Each of the enzymes is characterized by the reactions catalyzed. The hierarchy constructed using the EC numbering systems is called the enzyme hierarchy (e.g., [1.1.1.3] and [2.1.1]) which is called the EC class. When more than two enzymes are given as input, the EC class is the lowest class of all the upper classes of those enzymes in the enzyme hierarchy[19]. For example, [1.1] is the EC class between [1.1.1.3] and [1.1.2.4].

Here, reactions for which a part of the EC number is not specified, such as [1. − . − .−], and for which no EC numbers have yet been assigned are not in-

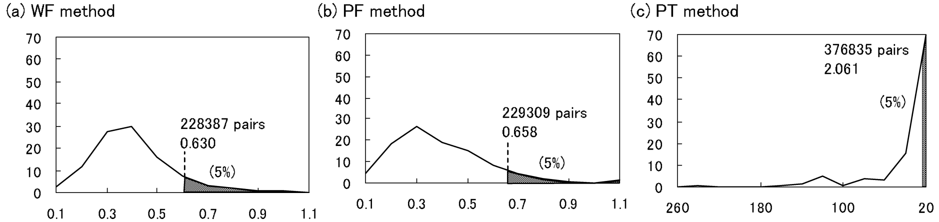


Fig. 3 Distributions of similarity scores. The scores were calculated using three methods: (a) whole fingerprint (WF) method, (b) partial fingerprint (PF) method, and (c) partial topological fragment spectra (PT) method.

cluded in the data. The 4,567,734 pairwise similarities of all of the reactions were calculated using the three types of reaction similarity measure: the whole fingerprint (WF) [5], the partial fingerprint (PF), and the partial topological fragment spectra (PT) methods.

In the WF method, the two components in Eq. (1) are replaced by the two compounds. The degrees of similarity $F'(c_1, c_2)$ of the bit-strings of compounds c_1 and c_2 are defined in accordance with the Tanimoto coefficient. The reaction similarity $WF(r_1, r_2)$ of the reaction (c_{11}, c_{12}) and the reaction (c_{21}, c_{22}) is calculated by the average of the compound similarities as follows:

$$WF(r_1, r_2) = (F'(c_{11}, c_{21}) + F'(c_{12}, c_{22}))/2 \qquad (8)$$

Figure 3 shows the distributions of pairwise similarities based on the different types of method, the Y-axis is the number of pairs and the X-axis is pairwise similarity. The top 5% reaction pairs from each distribution in Fig. 3 were extracted and classified according to their EC class level (**Table 1**). Hit number of reaction pairs in Table 1 indicates the number of reaction pairs in the top 5%. In Table 1, the first level of the EC class does not contain much information about the reaction mechanism. However, the third level, the EC sub-subclass, describes the type of bound or group on which the enzyme acts more specifically. Therefore, the ratio of the number of correctly detected reaction pairs in the third level was used as the true positive detection rate. Then, the PF and PT methods have an accuracy of 44.0% and 37.5%, respectively. These observations indicate that reaction similarities based on the PF method are better suited for agreement with the similarities based on the EC number than the PT method.

Table 1 The extent to which similarity according to EC agrees with similarity in terms of mechanism. EC level is the level up to which the two reactions share the same EC classification. Hit number of reaction pairs is the number of reaction pair in which two reactions are within the top 5% of whole reaction similarity. The percentages are shown in parentheses.

		Hit number of reaction pairs (%)					
EC level	# of pairs	WF		PF		PT	
≥ 0(all)	4567734	228387	(5.0)	229309	(5.0)	376835	(8.2)
≥ 1	1270206	73853	(5.8)	128403	(10.1)	205968	(16.2)
≥ 2	232100	29100	(12.5)	72871	(31.3)	66509	(28.6)
≥ 3	120184	19245	(16.0)	52896	(44.0)	45108	(37.5)
≥ 4	3681	2178	(59.1)	2349	(63.8)	2370	(64.3)

Table 2 Top 5 reaction similarities.

	WF				
Rank	Pair		Score	EC Pair	
1	R00010	R00026	1	[3.2.1.28]	[3.2.1.21]
2	R00010	R00028	1	[3.2.1.28]	[3.2.1.20]
3	R00010	R01678	1	[3.2.1.28]	[3.2.1.23]
4	R00010	R01718	1	[3.2.1.28]	[3.2.1.10]
5	R00010	R05549	1	[3.2.1.28]	[3.2.1.22]
	PF				
Rank	Pair		Score	EC Pair	
1	**R00006**	**R01841**	**1**	[2.2.1.6]	[4.1.2.30]
2	**R00006**	**R03403**	**1**	[2.2.1.6]	[4.1.2.30]
3	R00010	R00802	1	[3.2.1.28]	[3.2.1.48]
4	**R00010**	**R00803**	**1**	[3.2.1.28]	[2.4.1.7]
5	R00010	R00837	1	[3.2.1.28]	[3.2.1.93]
	PT				
Rank	Pair		Score	EC Pair	
1	R00005	R00472	0	[3.5.1.54]	[2.3.3.9]
2	R00005	R00776	0	[3.5.1.54]	[4.3.2.3]
3	R00006	R01841	0	[2.2.1.6]	[4.1.2.30]
4	R00006	R02948	0	[2.2.1.6]	[4.1.1.5]
5	R00006	R03403	0	[2.2.1.6]	[4.1.2.30]

The five most similar pairs of reactions for each method are shown in **Table 2** along with their EC numbers that agreed with each other. In the table, reaction pairs for which the EC number disagreed by the PF method are shown in bold. These five reactions are shown in **Fig. 4**. Here, reactions R00010 and R00803 differ at the class level of the EC ([3.2.1.28] and [2.4.1.7]). However, they are

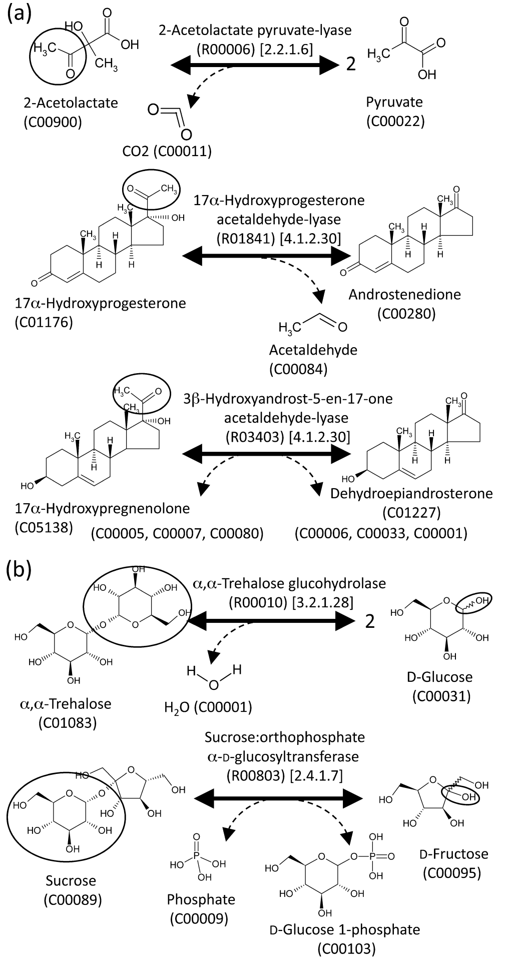


Fig. 4 Details of enzymatic reactions that differ at the class level of the EC number. The structural components modified by the enzymes are circled. Arrows shown in bold are main flows, and arrows shown in dotted lines are sub-flows based on KEGG. (a) Reactions of KEGG entries R00006, R01041, and R02403. (b) Reactions of KEGG entries R00010 and R00803.

identical according to the PF method, and the mechanisms of the reactions are very similar. These results suggest that the PF method would detect reaction similarities without being dependent on the EC number. Therefore, the PF method was used to align the metabolic pathways. Here, $S_{mean} = 0.335$ and $S_\sigma = 0.176$ (see Section 2.3).

3.2 Metabolic Pathway Alignment

Local alignments were performed between metabolic pathways in *E. coli* K-12 MG1655 using the fingerprint-based method targeting components. As the algorithm presented here does not consider branching pathways that occur in the metabolic network, a pre-processing procedure that extracts a set of non-branching sub-pathways is required. The metabolic pathway between the two compounds in the same metabolic map can be extracted using a shortest paths algorithm[20]. However, pathway reconstruction using a shortest paths algorithm has major problems caused by traversing irrelevant shortcuts through highly connected nodes, such as H_2O and ATP *etc.*[21]. To avoid this problem, the "reaction main" dataset in the KEGG database was used in this study. The major path data is represented by one adjacency matrix of a directed graph. A set of non-branching sub-pathways between any pair of compounds in the adjacency matrix were extracted using Dijkstra's standard shortest path algorithm[20] to simplify the extracting procedure. Target metabolic maps were limited to 37 (**Table 3**) to avoid duplication of extracted metabolic pathways. A total of 8,838 pathways were extracted (**Fig. 5**). Figure 5 shows the distribution of the extracted pathways against the distance for pairs of chemical compounds. Using the extracted pathways in *E. coli*, alignment was performed between any pairs of pathways classified into different metabolic maps.

Here, 2,000 random samplings were repeated 10 times, and the averages of the means and standard deviations of each set of alignments with pathway lengths less than 13 were calculated (**Fig. 6**). Using the empirical distribution of the means and standard deviation, each individual alignment score was converted into a *P*-value.

The top 20 alignments are shown in **Table 4**. This table shows *P*-values, alignment scores (Score), alignment results, and metabolic map IDs to which the input pathway belong. The alignment results are shown as a sequence of

Table 3 The 37 metabolic maps used in this analysis.

Map ID	Map Name	Map ID	Map Name
M00010	Glycolysis / Gluconeogenesis	M00400	Phenylalanine, tyrosine and tryptophan biosynthesis
M00020	Citrate cycle (TCA cycle)	M00450	Selenoamino acid metabolism
M00030	Pentose phosphate pathway	M00500	Starch and sucrose metabolism
M00040	Pentose and glucuronate interconversions	M00520	Nucleotide sugars metabolism
M00051	Fructose and mannose metabolism	M00530	Aminosugars metabolism
M00052	Galactose metabolism	M00561	Glycerolipid metabolism
M00130	Ubiquinone biosynthesis	M00620	Pyruvate metabolism
M00220	Urea cycle and metabolism of amino groups	M00630	Glyoxylate and dicarboxylate metabolism
M00230	Purine metabolism	M00640	Propanoate metabolism
M00240	Pyrimidine metabolism	M00650	Butanoate metabolism
M00251	Glutamate metabolism	M00670	One carbon pool by folate
M00252	Alanine and aspartate metabolism	M00710	Carbon fixation
M00260	Glycine, serine and threonine metabolism	M00730	Thiamine metabolism
M00271	Methionine metabolism	M00760	Nicotinate and nicotinamide metabolism
M00280	Valine, leucine and isoleucine degradation	M00770	Pantothenate and CoA biosynthesis
M00330	Arginine and proline metabolism	M00790	Folate biosynthesis
M00340	Histidine metabolism	M00860	Porphyrin and chlorophyll metabolism
M00360	Phenylalanine metabolism	M00910	Nitrogen metabolism
M00362	Benzoate degradation via hydroxylation		

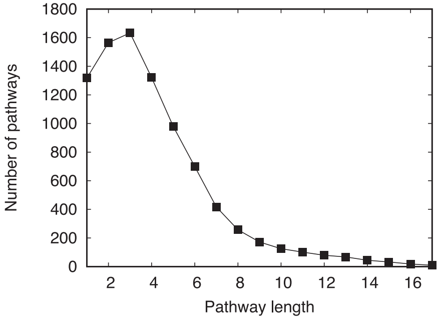


Fig. 5 Distribution of extracted pathways.

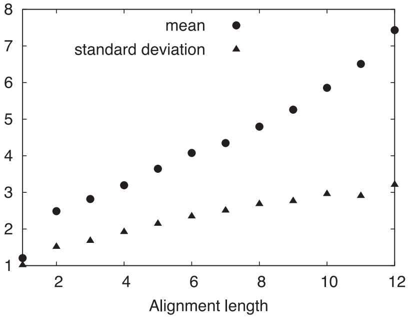


Fig. 6 Sampling results.

compound IDs representing its compounds and gap positions (indicated by "-"). The map IDs show the metabolic map category, corresponding to the formal names shown in Table 3. In these alignments, the highest score of the results is the score between purine and pyrimidine metabolism (map IDs M00230 and M00240). These two pathways consist of similar reaction sequence as shown in **Fig. 7** (a) (P-value = 4.59E-12). It is interesting that the structural changes in each compound in the two pathways are similar, for example both of them include a gap, and the final products are adenine and cytosine. Alignments of purine and pyrimidine metabolism often appear in the top 20 alignment results. Here we consider the 17th rank in Table 4, indicating alignment between fructose and mannose metabolism (M00051) and starch and sucrose metabolism (M00500). These two pathways consist of similar reaction sequences as shown in Fig. 7 (b)

Table 4 Top 20 alignments.

Rank	*P*-value	Score	Map ID	Alignment Results
1	4.59E-12	5.709	M00230	C00361 C00286 C00039 C00131 C00206 C00360 C00559 C00147
			M00240	C00105 C00015 C00046 C00112 ———— C00055 C00475 C00380
2	2.83E-11	5.000	M00230	C00360 C00559 ———— C00147 C00020 C00008 C00002
			M00240	C00239 C00881 C00526 C00106 C00105 C00015 C00075
3	2.83E-11	5.000	M00230	C00362 C00330 ———— C00242 C00144 C00035 C00044
			M00240	C00239 C00881 C00526 C00106 C00105 C00015 C00075
4	5.56E-11	4.962	M00230	C00286 C00330 ———— C00242 C00144 C00035 C00044
			M00240	C00239 C00881 C00526 C00106 C00105 C00015 C00075
5	1.57E-10	4.000	M00230	C00002 C00008 C00020 C00147 C00559
			M00240	C00075 C00015 C00105 C00106 C00526
6	1.57E-10	4.000	M00230	C00147 C00020 C00008 C00002 C00131
			M00240	C00106 C00105 C00015 C00075 C00460
7	1.57E-10	4.000	M00230	C00131 C00206 C00360 C00559 C00147
			M00240	C00459 C00363 C00364 C00214 C00178
8	1.57E-10	4.000	M00230	C00330 C00242 C00144 C00035 C00044
			M00240	C00526 C00106 C00105 C00015 C00075
9	1.57E-10	4.000	M00230	C00387 C00144 C00035 C00044 C00286
			M00240	C00475 C00055 C00112 C00063 C00458
10	1.57E-10	4.000	M00230	C00387 C00144 C00035 C00044 C00286
			M00240	C00299 C00105 C00015 C00075 C00460
11	1.57E-10	4.000	M00230	C00242 C00144 C00035 C00044 C00286
			M00240	C00106 C00105 C00015 C00075 C00460
12	5.77E-10	3.934	M00230	C00212 C00147 C00020 C00008 C00002
			M00240	C00526 C00106 C00105 C00015 C00075
13	5.77E-10	3.934	M00230	C00131 C00206 C00360 C00559 C00147
			M00240	C00063 C00112 C00055 C00475 C00380
14	1.20E-09	4.775	M00230	C00361 C00286 C00039 C00131 C00206 C00360 C00559
			M00240	C00105 C00015 C00046 C00112 ———— C00055 C00475
15	1.63E-09	5.331	M00230	C00362 C00361 C00286 C00039 C00131 C00206 C00360 C00559
			M00240	C00299 C00105 C00015 C00046 C00112 ———— C00055 C00475
16	1.67E-09	4.754	M00230	C00286 C00039 C00131 C00206 C00360 C00559 C00147
			M00240	C00075 C00046 C00112 C00055 ———— C00475 C00380
17	2.48E-09	3.857	M00051	C00159 C00275 C00636 C00096 C02492
			M00500	C00267 C00668 C00103 C00029 C00760
18	2.56E-09	4.727	M00230	C00212 C00147 C00020 C00008 ———— C00002 C00131
			M00240	C00526 C00106 C00105 C00015 C00046 C00112 C00705
19	3.40E-09	4.709	M00230	C00286 C00039 C00131 C00206 C00360 C00559 C00147
			M00240	C00015 C00046 C00112 ———— C00055 C00475 C00380
20	5.82E-09	5.775	M00230	C00361 C00286 C00039 C00131 ———— C00206 ———— C00360 C00559 C00147
			M00240	C00055 C00112 C00046 C00015 C01346 C00365 C00364 ———— C00214 C00178

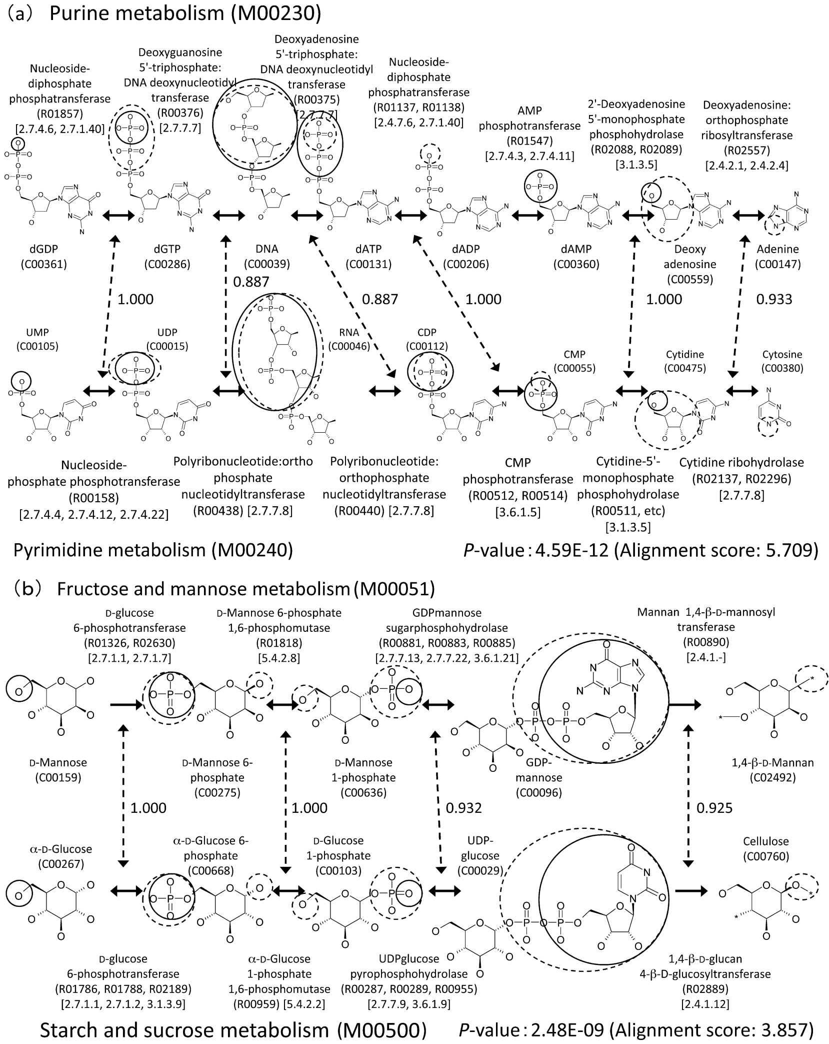

(a) Purine metabolism (M00230)
Nucleoside-diphosphate phosphatransferase (R01857) [2.7.4.6, 2.7.1.40]
Deoxyguanosine 5'-triphosphate: DNA deoxynucleotidyl transferase (R00376) [2.7.7.7]
Deoxyadenosine 5'-triphosphate: DNA deoxynucleotidyl transferase (R00375) [2.7.7.7]
Nucleoside-diphosphate phosphatransferase (R01137, R01138) [2.4.7.6, 2.7.1.40]
AMP phosphotransferase (R01547) [2.7.4.3, 2.7.4.11]
2'-Deoxyadenosine 5'-monophosphate phosphohydrolase (R02088, R02089) [3.1.3.5]
Deoxyadenosine: orthophosphate ribosyltransferase (R02557) [2.4.2.1, 2.4.2.4]
dGDP (C00361)
dGTP (C00286)
DNA (C00039)
dATP (C00131)
dADP (C00206)
dAMP (C00360)
Deoxy adenosine (C00559)
Adenine (C00147)
1.000
0.887
0.887
1.000
1.000
0.933
UMP (C00105)
UDP (C00015)
RNA (C00046)
CDP (C00112)
CMP (C00055)
Cytidine (C00475)
Cytosine (C00380)
Nucleoside-phosphate phosphotransferase (R00158) [2.7.4.4, 2.7.4.12, 2.7.4.22]
Polyribonucleotide:orthophosphate nucleotidyltransferase (R00438) [2.7.7.8]
Polyribonucleotide: orthophosphate nucleotidyltransferase (R00440) [2.7.7.8]
CMP phosphotransferase (R00512, R00514) [3.6.1.5]
Cytidine-5'-monophosphate phosphohydrolase (R00511, etc) [3.1.3.5]
Cytidine ribohydrolase (R02137, R02296) [2.7.7.8]
Pyrimidine metabolism (M00240)
P-value : 4.59E-12 (Alignment score: 5.709)
(b) Fructose and mannose metabolism (M00051)
D-glucose 6-phosphotransferase (R01326, R02630) [2.7.1.1, 2.7.1.7]
D-Mannose 6-phosphate 1,6-phosphomutase (R01818) [5.4.2.8]
GDPmannose sugarphosphohydrolase (R00881, R00883, R00885) [2.7.7.13, 2.7.7.22, 3.6.1.21]
Mannan 1,4-β-D-mannosyl transferase (R00890) [2.4.1.-]
D-Mannose (C00159)
D-Mannose 6-phosphate (C00275)
D-Mannose 1-phosphate (C00636)
GDP-mannose (C00096)
1,4-β-D-Mannan (C02492)
1.000
1.000
0.932
0.925
α-D-Glucose (C00267)
α-D-Glucose 6-phosphate (C00668)
D-Glucose 1-phosphate (C00103)
UDP-glucose (C00029)
Cellulose (C00760)
D-glucose 6-phosphotransferase (R01786, R01788, R02189) [2.7.1.1, 2.7.1.2, 3.1.3.9]
α-D-Glucose 1-phosphate 1,6-phosphomutase (R00959) [5.4.2.2]
UDPglucose pyrophosphohydrolase (R00287, R00289, R00955) [2.7.7.9, 3.6.1.9]
1,4-β-D-glucan 4-β-D-glucosyltransferase (R02889) [2.4.1.12]
Starch and sucrose metabolism (M00500)
P-value : 2.48E-09 (Alignment score: 3.857)

Fig. 7 Details of alignment results. Each enzyme name is shown above the reaction IDs and EC numbers in KEGG, and their substrates and products are shown between the enzymes. The combinations of reaction correspondence determined by alignment are connected with dashed arrows, and the similarity scores of the reaction are shown horizontally. Components are circled and corresponded by solid and dashed lines.

(P-value = 2.48E-09). Mannan 1,4-β-D-mannosyl transferase (R00890) has not been assigned a complete EC number ([2.4.1.−]), which corresponded to 1,4-β-D-glucan 4-β-D-glucosyltransferase (R02889) with the EC number [2.4.1.12].

4. Conclusions and Future Work

Modified structural components were extracted from compound structures constituting the enzymatic reaction using RPAIR data in KEGG. Reaction similarity measures based on extracted compound structures were proposed, and these scoring systems were evaluated in comparison to EC classification based scoring systems. The method was applied to align the metabolic pathways in *E. coli*. The alignment results extracted two groups of metabolic pathway: purine and pyrimidine metabolism, fructose and mannose metabolism, and starch and sucrose metabolism. These results suggested that the proposed method would detect pathway similarities without being dependent on the EC number. However, the two reaction similarity measures described here should be regarded as starting points for the development of similarity measures for reaction mechanisms. In future studies, it will be necessary to evaluate the measures in more detail, improve the pathway extraction algorithm, and classify the alignment results.

References

1) Dandekar, T., Schuster, S., Snel, B., Huynen, M. and Bork, P.: Pathway alignment: application to the comparative analysis of glycolytic enzymes, *Biochemical J.*, Vol.343, No.1, pp.115–124 (1999).

2) Galperin, M.Y., Walker, D.R. and Koonin, E.V.: Analogous enzymes: independent inventions in enzyme evolution, *Genome Res.*, Vol.8, No.8, pp.779–790 (1998).

3) Webb, E.C. (Eds): *Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*, Academic Press (1993).

4) O'Boyle, N.M., Holliday, G.L., Almonacid, D.E. and Mitchell, J.B.O.: Using reaction mechanism to measure enzyme similarity, *J. Mol. Biol.*, Vol.368, No.5, pp.1484–1499 (2007).

5) Tohsato, Y. and Nishimura, Y: Metabolic pathway alignment based on similarity between chemical structures, *IPSJ Transactions on Bioinformatics*, Vol.48, No.SIG17(TBIO3), pp.736–745 (2007).

6) Garey, M.R. and Johnson, D.S.: *Computers and intractability: a guide to the theory of NP-completeness*, Freeman and Company (1979).

7) Hattori, M., Okuno, Y., Goto, S. and Kanehisa, M.: Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways, *J. Am. Chem. Soc.*, Vol.125, No.39, pp.11853–11865 (2003).

8) Xue, L., Godden, J.W., Stahura, F.L. and Bajorath, J.: Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys, *J. Chem. Inf. & Comput. Sci.*, Vol.43, No.4, pp.1218–1225 (2003).

9) Kotera, M., Okuno, Y., Hattori, M., Goto, S. and Kanehisa, M.: Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions, *J. Am. Chem. Soc.*, Vol.126, pp.16487–16498 (2004).

10) Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. and Yamanishi, Y.: KEGG for linking genomes to life and the environment, *Nucleic Acids Res.*, Vol.36, pp.D480–D484 (2008).

11) Teichmann, S.A., Rison, S.C., Thornton, J.M., Riley, M., Gough, J. and Chothia, C.: The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli, *J. Mol. Biol.*, Vol.311, No.4, pp.683–708 (2001).

12) Schmidt, S., Sunyaev, S., Bork, P. and Dandekar, T.: Metabolites: a helping hand for pathway evolution?, *Trends in Biochemical Sciences*, Vol.28, No.6, pp.336–341 (2003).

13) Takahashi, Y., Ohoka, H. and Ishiyama, Y.: Structural similarity analysis based on topological fragment spectra, *Advances in Molecular Similarity*, Vol.2, pp.93–104 (1998).

14) Takahashi, Y., Konji, M. and Fujishima S.: MolSpace, a computer desktop tool for visualization of massive molecular data, *J. Mol. Graph. Model.*, Vol.21, pp.333–339 (2003).

15) MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.

16) Smith, T.F. and Waterman, M.S.: Identification of common molecular subsequences, *J. Mol. Biol.*, Vol.147, pp.195–197 (1981).

17) Open Babel: http://openbabel.org/

18) MacCuish, N.E. and MacCuish, J.D.: Clustering compound data: asymmetric clustering of chemical datasets, chemometrics and cheminfomatics, *ACS Symposium Series*, Vol.894, ed. B.K. Lavine, Oxford University Press, (2005). http://www.mesaac.com/

19) Tohsato, Y., Matsuda, H. and Hashimoto, A.: An application of a pathways alignment method to the analysis of metabolic pathways, *Res. Comm. In Biochem., Cell & Mol. Biol.*, Vol.5, pp.179–191 (2003).

20) Arita, M.: In silico atomic tracing by substrate-product relationships in Escherichia coli intermediary metabolism, *Genome Res.*, Vol.13, No.11, pp.2455–2466 (2003).

21) Arita, M.: The metabolic world of Escherichia coli is not small, *Proc. Natl. Acad. Sc. U.S.A*, Vol.101, No.6, pp.1543–1547 (2004).

(Received August 31, 2008)
(Accepted September 30, 2008)
(Released March 24, 2009)

(Communicated by *Yoshihiro Taguchi*)

Yukako Tohsato is an Assistant Professor at the Department of Bioscience and Bioinformatics, Ritsumeikan University. She received her M.E. degree from Kyushu Institute of Technology University in 1997. She worked at Mitsubishi Electric Co. from 1997 to 1999. She received her Ph.D. from Osaka University in 2002. From 2002 to 2003, she worked as a Research Associate at Osaka University. From 2003 to 2004, she worked as a Researcher at Osaka University. She is a member of IPSJ.

Yu Nishimura received B.Sc. from Ritsumeikan University in 2007. Since 2007, he has been a Masters course student at the Graduate School of Science and Engineering, Information Science & Systems Engineering Major, Bioinformatics Science Course at Ritsumeikan University.