Original Paper

Metabolic Pathway Alignment Based on Similarity between Chemical Structures

Yukako Tohsato† and Yu Nishimura††

In many of the chemical reactions in living cells, enzymes act as catalysts in the conversion of certain compounds (substrates) into other compounds (products). Metabolic pathways are formed as the products of these reactions are used as the substrates of other reactions. Comparative analyses of the metabolic pathways among species provide important information on both evolution and potential pharmacological targets. Here, we propose a method to align the metabolic pathways based on similarities between chemical structures. To measure the degree of chemical similarity, we formalized a scoring system using the MACCS keys and the Tanimoto/Jaccard coefficients. To determine the effectiveness of our method, it was applied to analyses of metabolic pathways in *Escherichia coli*. The results revealed compound similarities between fructose and mannose biosynthesis and galactose biosynthesis pathways.

1. Introduction

To obtain the energy necessary for cellular activities, *in vivo* cells take up many types of material in the form of food. The cells break down and synthesize materials required for self-maintenance and growth *via* an enormous number of chemical reactions. These chemical reactions occurring in an organism are known collectively as "metabolism," which consists of enzymatic reactions that result in the conversion of certain compounds (substrates) into other compounds (products) by virtue of the action of enzymes (proteins). A large-scale and complex metabolic network is formed as the products of these reactions feed into other reactions as substrates. Data regarding these reactions are now available in several public databases, including KEGG [1] and MetaCyc [2] *via* the World Wide Web.

For example, KEGG is a collection of manually drawn metabolic maps. The pathway of a reaction is generally called the "metabolic pathway," which is often abbreviated simply to "pathway." Comparative analysis of the metabolic pathways among different species provides essential information on both the evolution of organisms and on potential pharmacological targets, and there has been a great deal of research in this area in recent years [3].

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

We focused on "pathway duplication," which is one of the hypotheses explaining the evolution of the metabolic networks. Pathway duplication suggests that evolution has occurred through duplication of the genes encoding proteins within a pathway [4]. This example of pathway duplication was described by Huynen & Snel, who also noted pathway duplication of proteins in the *prp* operon and of those in the glyoxylate shunt [5]. Tryptophan and histidine biosynthesis are two pathways with similar reaction chemistries that are catalyzed by homologous enzymes, and may the result of pathway duplication.

Pathway duplication is usually detected 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 [6]. Therefore, emphasis is placed on comparison and investigation analysis results from a variety of standpoints. If the structures of the proteins being compared are known, both their domain structures and evolutionary relationships, and the combinations of structural, functional, and sequence information are using for detection of pathway duplication [5].

In our previous study in which we aligned pathways based on the functional hierarchy of enzymes using EC numbers [7], we found that compounds were converted in similar ways when two sequences of EC numbers were similar [8]. However, the EC numbers represent artificial classifications. There is a large degree

of deviation in distribution among the enzyme hierarchy. There are enzymatic reactions for which no EC numbers have yet been assigned. For example, although *Escherichia coli* K-12 MG1655 has 1115 enzyme reactions, 123 are not labeled by EC number, including those for which a part of the EC number is not specified, such as $[1.-.-.-]$, and these enzyme reactions represent about 11% of the total.

Here, we focus on the structural formulae of compounds ("compound structures"). As finding the exact similarity between two compound structures is an NP hard problem, a number of methods for measuring the similarity between compound structures have been proposed. The main approach is the fingerprint-based comparison [9], 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 [9]. In the present study, we defined the similarity between two compound structures based on their descriptors, and we propose a method for alignment between metabolic pathways based on the similarity. Here, we report the results obtained by applying the proposed method to actual metabolic pathway data in *E. coli* K-12 MG1655.

2. Methods

2.1 Representation Framework

A metabolic network is a set of biochemical reactions, *i.e.*, reactions that convert one or more compounds into one or more other compounds. The network is modeled by an undirected graph, $G = (C, R)$, where C is the set of nodes representing the compounds, R is a set of undirected link-pairs of one compound node in C and one reaction node in R. When reaction $r \in R$ converts a certain compound (substrate) $c_1 \in C$ to another compound (product) $c_2 \in C$, the reaction is represented as $c_1 \xrightarrow{r} c_2$. Reversible reactions, such as $c_1 \xleftrightarrow{r} c_2$, are separated into two reactions, $c_1 \xrightarrow{r} c_2$ and $c_2 \xrightarrow{r} c_1$. The representation $c_1 \xrightarrow{r} c_2$ is the same as a pair of compounds (c_1, c_2). Two reactions are *adjacent* if there exists an edge r, *i.e.*, $(c_1, c_2) \in R$, the edge set of this graph, if they share at least one chemical compound as either as substrate or as product. In present study, a metabolic pathway from metabolite c_1 to c_m is definedas a sequence $(c_1, c_2)(c_2, c_3) \cdots (c_{m-1}, c_m)$ of pairs of compounds, and is identified with a sequence $r_1 r_2 \cdots r_m$ of biochemical reactions that are adjacent to each other. The length of the pathway is the number of pairs of compounds.

2.2 Similarity Score between Reactions

The similarity between reactions is defined with reference to similarity between the chemical structures of the pair of compounds involved. Some descriptors have been proposed to facilitate treatment of the structures of compounds on a computer. The descriptors currently in wide use are fingerprints, which show whether the specific molecular structure would exist in a molecule, and such an expressions are designed for database indexing to increase speed in substructure searches.

We use MACCS keys [10], which are some of the most widely used descriptors and are encoded in 166 bits. The keys are bit-strings, typically showing the presence (bit = 1) or absence (bit = 0) of a predefined structural element (see Appendix for details). The bit-string of compound c is represented by $B(c) = (x_1, x_2, \cdots x_n)$. The keys are 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 define the degree of similarity, a variety of numerical methods have been proposed [9]. In the present study, we used the Tanimoto (Jaccard) coefficient, which is an index of the relative correlations between two bit-strings. The degrees of similarity $T(c_1, c_2)$ of the bit-strings $B(c_1)$ and $B(c_2)$ are defined in accordance with the Tanimoto coefficient as follows:

$$T(c_1, c_2) = \frac{B(c_1) \cap B(c_2)}{B(c_1) \cup B(c_2)} \quad (1)$$

$B(c_1) \cap B(c_2)$ is the number of common 1 bits in both bit-strings $B(c_1)$ and $B(c_2)$. $B(c_1) \cup B(c_2)$ is the number of 1 bits in either bit-strings $B(c_1)$ and $B(c_2)$. By definition, $T(c_1, c_2)$ is the number in the range 0 to 1; the closer to 1, the higher the degree of similarity between the two bit-strings, while the closer to 0, the lower the degree of similarity between the two bit-strings.

The reaction similarity $S(r_1, r_2)$ of the reaction $c_{11} \xrightarrow{r_1} c_{12}$ and the reaction $c_{21} \xrightarrow{r_2} c_{22}$ as shown in **Fig. 1** is calculated by the average of the compound similarities of corresponding compounds as follows:

$$S(r_1, r_2) = \frac{T(c_{11}, c_{21}) + T(c_{12}, c_{22})}{2} \quad (2)$$

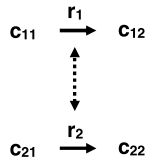

Fig. 1 Relation of combination of reactions.

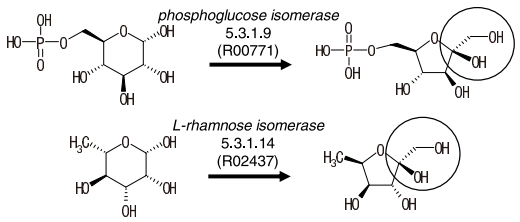

Fig. 2 Example of combination of similar reactions.

To calculate the reaction similarity, it is desirable to compare the differences in the bit-strings $B(c_{11})$ and $B(c_{12})$ and in the bit-strings $B(c_{21})$ and $B(c_{22})$. However, the number of changed bits often differs greatly in enzyme reactions that cause structural changes in similar compounds. For example, changes in the compound structure catalyzed by phosphoglucose isomerase and L-rhamnose isomerase are similar (**Fig. 2**). In this figure, the structural components modified by the enzymes are circled and the enzyme names are shown above their EC numbers (KEGG reaction IDs are given in parentheses). In this case, the number of the bits changed by phosphoglucose isomerase is 1, the number of the bits changed by L-rhamnose isomerase is 13, and the number of the bits that changed in common is 1. This is considered to be a phenomenon to cut as the classification of MACCS key allows duplication of partial structures. Therefore, in this research we compared the bit-strings of substrates with those of products, which is close to the original use of MACCS keys.

2.3 Metabolic Pathway Alignment Algorithm

In this study, we extended the local alignment algorithm based on dynamic programming (Smith and Waterman algorithm)[11]. 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:

procedure *PathwayAlignment*
input: $p_1 = r_{11}r_{12}\ldots r_{1m}$, $p_2 = r_{21}r_{22}\ldots r_{2n}$;
output: Optimal local alignment, alignment score;
for $i := 0$ **to** m **do** $M[i, 0] := 0$;
for $j := 0$ **to** n **do** $M[0, j] := 0$;
$max := 0$; $i_{max} := 0$; $j_{max} := 0$;
for $i := 1$ **to** m **do**
  for $j := 1$ **to** n **do**
    $M[i, j] :=$
    $\max \begin{cases} 0, \\ M[i-1, j-1] + S(r_i, r_j) \\ M[i, j-1] + d, \\ M[i-1, j] + d, \end{cases}$
    if $max > L[i, j]$ **then**
      $max := L[i, j]$; $i_{max} := i$; $j_{max} := i$;
$i := i_{max}$; $j := j_{max}$;
$align\_p_1 =$ “”; $align\_p_2 =$ “”;
while $(M[i, j]! = 0)$
  if $M[i, j] = M[i-1, j-1] + S(r_{1i}, r_{2j})$ **then**
    push r_{1i} to $align\_p_1$;
    push r_{2j} to $align\_p_2$;
    $i := i - 1$;
    $j := j - 1$;
  else if $M[i, j] = M[i-1, j] + d$ **then**
    push r_{1i} to $align\_p_1$;
    push “-” to $align\_p_2$;
    $i := i - 1$;
  else $M[i, j] = M[i, j-1] + d$
    push “−” to $align\_p_1$;
    push r_{2j} to $align\_p_2$;
    $j := j - 1$;
endwhile
return $align\_p_1$, $align\_p_2$, max;

Fig. 3 Pathway alignment algorithm.

$$M[i, j] = \max \begin{cases} 0, \\ M[i-1, j-1] + S(r_i, r_j) \\ M[i, j-1] + d, \\ M[i-1, j] + d, \end{cases} \tag{3}$$

Thus, we extended the alignment algorithm 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, and we set the gap penalty to -1.

Once filling of the matrix is completed, the score of optimal local alignment of the sequences is the highest score. The traceback procedure starts at the highest value of $M(i, j)$ over the whole matrix, and ends when it reaches a cell with a value of 0, which corresponds to the optimal alignment. The algorithm is shown in **Fig. 3**. The running-time complexity of the alignment is $O(mn)$ where m and n are the maximum lengths of the two pathways given as input.

The alignment score was divided into the length of the alignment to avoid the influence of the sequence size. We extracted alignment results with corrective scores above a given threshold value.

3. Experiments and Results

3.1 Experimental Data

We performed local alignments between metabolic pathways in *E. coli* K-12 MG1655 obtained from the KEGG database (Version 40.0, 2006/10). Bit-string data were generated from SMILIES data in the PubChem database using the Fingerprint Module of MESA in OpenEye Scientific Software [12], which generated 164 bit-strings from SMILES strings input [13]. The bit-strings are a public subset of 166 MACCS keys. These generated data were linked to KEGG using cross-references to PubChem.

As our algorithm 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 shortest paths algorithms (e.g., Dijkstra's algorithm) [14]. Dijkstra's algorithm finds the shortest path from the start-compound to all reachable compounds connected to it in the same network (in this case, one network coresponds to one metabolic map). However, pathway reconstruction using a shortest paths algorithm has major problems caused by traversing irrelevant shortcuts through highly connected nodes, which correspond to metabolites and co-factors (e.g., H_2O and ATP) [14].

Therefore, in this study, we used major path data categorizing "reaction_main" in KEGG. A network in the major path data is represented by an adjacency matrix which defines an unweighted graph as shown in **Fig. 4**. We extracted a set of non-branching sub-pathways between any pair of compounds in the adjacency matrix using Dijkstra's algorithm. Target metabolic maps are limited to 37 (**Table 1**) to avoid duplication of extracted metabolic pathways. As a result, we obtained 15,444 extracted pathways of lengths greater than two.

3.2 Alignment Results and Discussion

Using the 15,444 extracted pathways in *E. coli* (details described in Section 3.1), we performed alignments between any pair of the pathways classified into different metabolic maps. The alignment score was divided into the length of the alignment to avoid the influence of sequence size. We extracted alignment results with corrective scores greater than a given threshold value. The extracted alignments were classified according to the alignment length and scored in order of decreasing alignment score.

Setting the alignment threshold to 0.95, we obtained longest alignment of length 5 in the alignment results. All alignment results in alignments of lengths 5 and 4 are shown in **Table 2**; this table shows alignment scores,

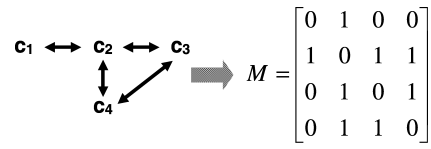

Fig. 4 Metabolic map and its adjacency matrix.

Table 1 The 37 metabolic maps used in this analysis.

Map ID	Map Name
MAP00010	Glycolysis / Gluconeogenesis
MAP00020	Citrate cycle (TCA cycle)
MAP00030	Pentose phosphate pathway
MAP00040	Pentose and glucuronate interconversions
MAP00051	Fructose and mannose metabolism
MAP00052	Galactose metabolism
MAP00130	Ubiquinone biosynthesis
MAP00220	Urea cycle and metabolism of amino groups
MAP00230	Purine metabolism
MAP00240	Pyrimidine metabolism
MAP00251	Glutamate metabolism
MAP00252	Alanine and aspartate metabolism
MAP00260	Glycine, serine and threonine metabolism
MAP00271	Methionine metabolism
MAP00280	Valine, leucine and isoleucine degradation
MAP00330	Arginine and proline metabolism
MAP00340	Histidine metabolism
MAP00360	Phenylalanine metabolism
MAP00362	Benzoate degradation via hydroxylation
MAP00400	Phenylalanine, tyrosine and tryptophan biosynthesis
MAP00450	Selenoamino acid metabolism
MAP00500	Starch and sucrose metabolism
MAP00520	Nucleotide sugars metabolism
MAP00530	Aminosugars metabolism
MAP00561	Glycerolipid metabolism
MAP00620	Pyruvate metabolism
MAP00630	Glyoxylate and dicarboxylate metabolism
MAP00640	Propanoate metabolism
MAP00650	Butanoate metabolism
MAP00670	One carbon pool by folate
MAP00710	Carbon fixation
MAP00730	Thiamine metabolism
MAP00760	Nicotinate and nicotinamide metabolism
MAP00770	Pantothenate and CoA biosynthesis
MAP00790	Folate biosynthesis
MAP00860	Porphyrin and chlorophyll metabolism
MAP00910	Nitrogen metabolism

Table 2 Alignment results.

(a) Alignment results of length 5

Score	Map ID	Alignment
0.968	MAP00051	C01131 C00111 C05378 C05345 C00644 C00392
	MAP00052	C01286 C00118 C03785 C01097 C06311 C01697

(b) Alignment results of length 4

Score	Map ID	Alignment
0.992	MAP00051	C00111 C05378 C05345 C00644 C00392
	MAP00052	C00118 C03785 C01097 C06311 C01697
0.992	MAP00051	C00118 C05378 C05345 C00644 C00392
	MAP00052	C00111 C03785 C01097 C06311 C01697
0.977	MAP00052	C00103 C00446 C00124 C00243 C05396
	MAP00500	C05345 C00668 C00267 C00208 C02995
0.960	MAP00051	C01131 C00111 C05378 C05345 C00644
	MAP00052	C01286 C00118 C03785 C01097 C06311
0.957	MAP00051	C01094 C05378 C05345 C00644 C00392
	MAP00052	C00118 C03785 C01097 C06311 C01697
0.957	MAP00051	C01094 C05378 C05345 C00644 C00392
	MAP00052	C00111 C03785 C01097 C06311 C01697
0.955	MAP00052	C00052 C00446 C00124 C05402 C00031
	MAP00500	C00029 C00103 C00089 C00267 C00208
0.955	MAP00052	C00052 C00446 C00124 C05400 C00159
	MAP00500	C00029 C00103 C00089 C00267 C00208

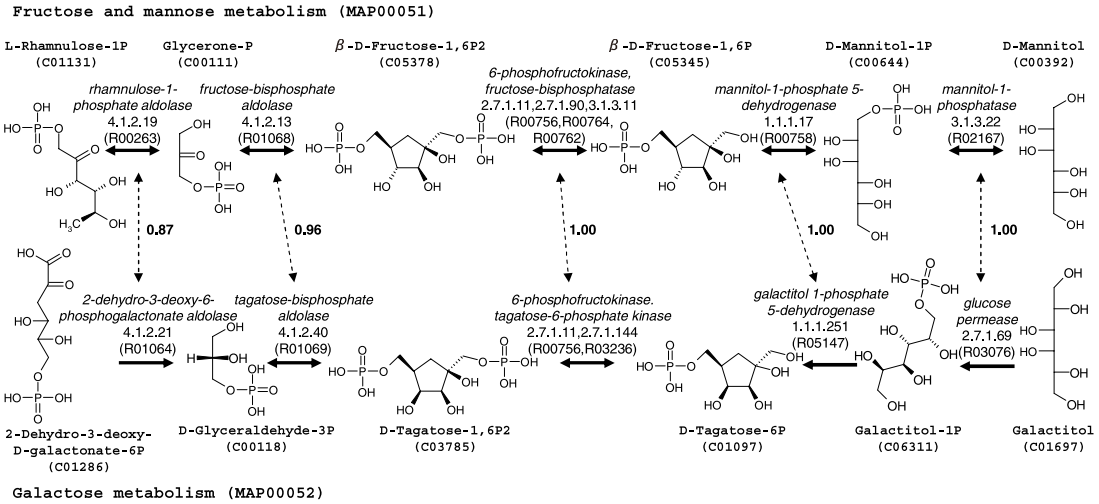

Fig. 5 Top alignment result with alignment length 5.

alignment results, and metabolic map IDs to which the input pathways belong. The alignment scores are already divided with alignment length. The alignment results are shown as a sequence of compound IDs representing its compounds. The map ID shows the category of metabolic map, corresponding to their formal names shown in Table 1.

In these alignments, the highest score of the results in the case of length 5 is the score between fructose and mannose metabolic pathways and galactose metabolic pathways. These two pathways consist of similar reaction sequences as shown in **Fig. 5**. Each enzyme name is shown above the EC numbers and reaction IDs in KEGG, and their substrates and products are shown between the enzymes. The trends of the arrows of reactions express the flow of the reactions that may actually occur. The combinations of reaction correspondence determned by alignment are connected with dashed arrows, and the similarity scores of the reaction are shown horizontally.

It is interesting that the structural changes in each compound in the two pathways are similar. Moreover, the portion that could be aligned overlapped in the alignment result based on EC number reported in the litera-

ture (overlapping parts of EC numbers occur in ([4.1.2.19], [4.1.2.21]), ([4.1.2.13], [4.1.2.40]), and ([2.7.1.11], [2.7.1.11])) [8]. Between the reactions to which the EC numbers [3.1.3.22] and [2.7.1.69] were attached, although EC numbers differ, the structural changes in each compound in the two reactions are similar.

Figure 6 shows the details of changes in the compound structure in the alignment results with alignment length 4. In these alignments, the structural changes in each compound in the two pathways are also similar, as in Fig. 5.

3.3 Comparison of Two Types of Similarity Score

In this section, we calculated the correlation between proposed similarity score and similarity score based on the EC numbering system [8].

Each of the enzymes is characterized by the reactions catalyzed. The International Union of Biochemistry and Molecular Biology (IUBMB) developed a classification scheme based on this observation. The hierarchy constructed using the EC numbering system is called the enzyme hierarchy. In addition, each element of the enzyme hierarchy (e.g., [1.1.1.3], [2.1.1], and [∗]) is called an enzyme class ([*] is the top level of the enzyme class, which expresses arbitrary enzymes). Given more than two enzymes as input, the enzyme class, which is the lowest class of all the upper classes of those enzymes in the enzyme hierarchy, is called the *common upper class*. Between the same enzymes, their common upper class is the same as their enzyme class. For example, [1.1] is the common upper class between [1.1.1.3] and [1.1.2.4] (details shown in Ref. 8)). Then, the similarity score based on the enzyme hierarchy between two reactions r_1, r_2 is calculated by the following formula:

$$E(r_1, r_2) = -\log_2 \frac{n}{m} \qquad (4)$$

where n is the number of reactions including the common upper class between r_1 and r_2, and the constant m is the number of all reactions.

In this paper, we set $m = 1061$. Although the above value of the information content is dependent on the given pathway set, the information content of the highest-level class is always zero as $n/m = 1$. Moreover, for any enzyme class h, $I(h) \leq I(h_i)$ $(1 \leq i \leq n)$ holds where $h_1, h_2, \cdots h_n$ are sub-classes of h. In the case of a reaction with different EC numbers, we select the combination that yields the highest score.

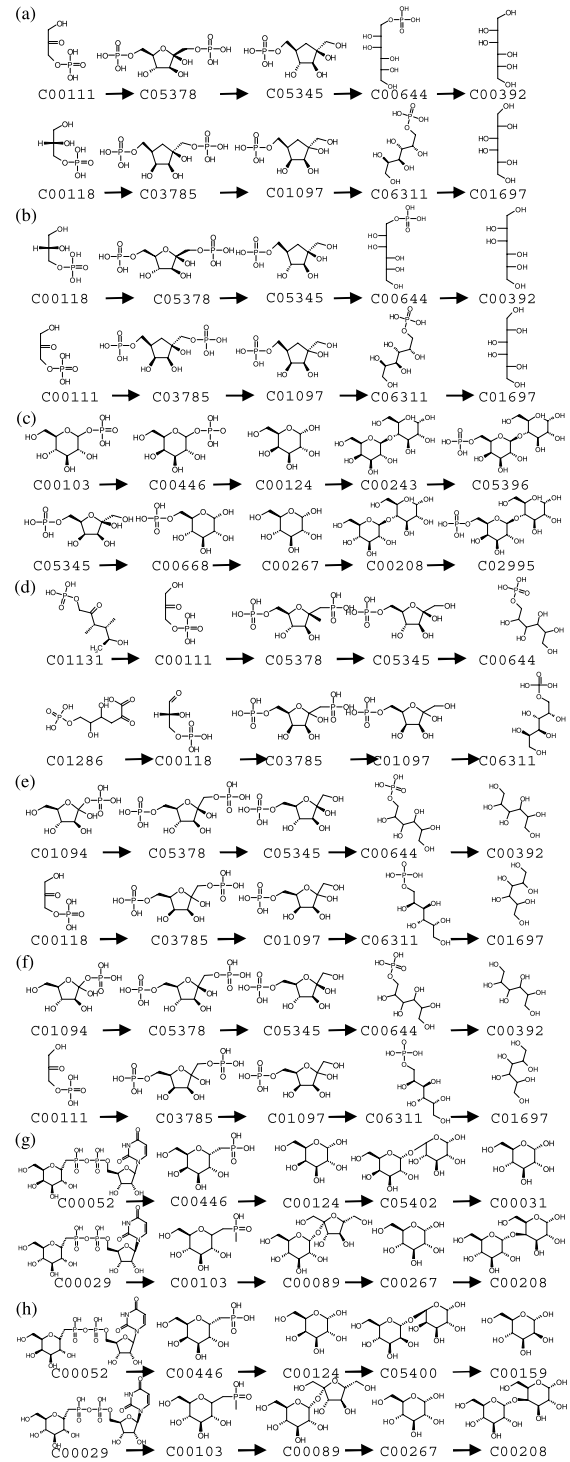

Fig. 6 Top alignment results with alignment length 4.

When there are two or more compounds that are the substrate and product in an enzymatic reaction exist, the similarity score between the two reactions differs in score according to which

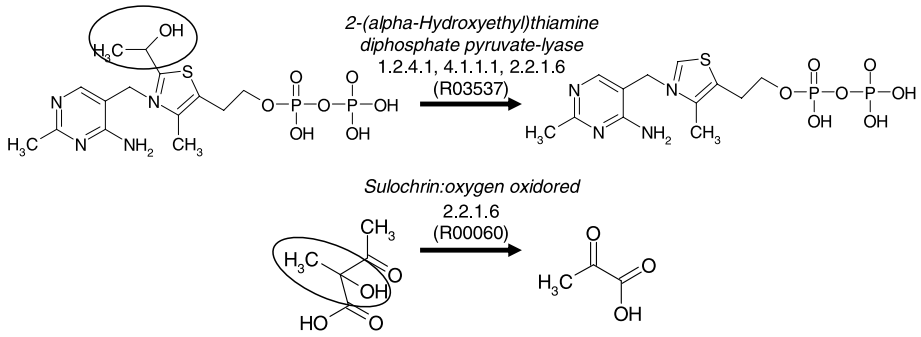

Fig. 8 Example of a combination of reactions.

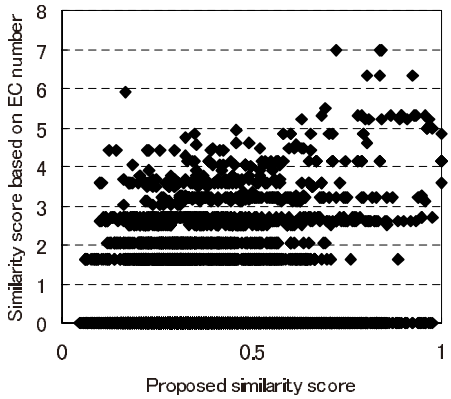

Fig. 7 Correlation of the proposed similarity score and the similarity score based on EC number.

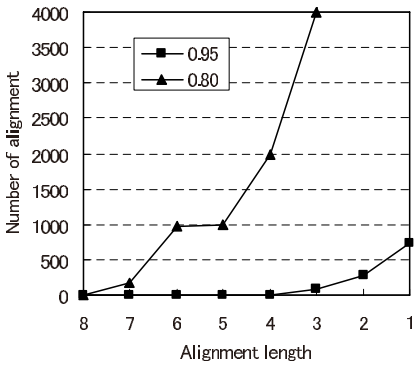

Fig. 9 The number of alignments according to alignment length.

substrate and product combination is chosen. Since the path is not specified when calculating for correlation, the combination cannot be decided. Therefore, the similarity score was chosen as the combination yielding the highest score.

The results are plotted in **Fig. 7** against similarity score $S(r_1, r_2)$ and similarity score $E(r_1, r_2)$ between pairs of all reactions in 37 metabolic maps in Table 1. The similarity score based on EC number in the case of combination of the same reactions on the horizontal and vertical axes for the proposed score was determined.

Although the scores obtained using the proposed method are high based on EC number although they are low according to that based on EC number, many proposed scores are looked at by the figure rather than the phenomenon which becomes low. The proposed scoring method is considered to perform finer matching than that based on EC number.

However, although the similarity score based on EC number is low, the combination of reactions yielded a high score based on the proposed similarly score. One of the combinations of reactions is shown in **Fig. 8**. This is the combination of the reaction from which the score based on EC number was set to 5.900, and the proposed similarity score was set to 0.167 in Fig. 7. In Fig. 8, the structural components that are modified by the enzymes are circled. The combination of EC numbers in which the score based on EC number becomes highest is [2.2.1.6] and [2.2.1.6]. This figure confirms that the combinations of compounds that constitute each reaction are completely different. Thus, the scoring system requires improvement.

3.4 Alignment Score

Finally, we examined the influence of compensation of the alignment score. The number of alignments changes with the threshold. **Figure 9** shows the changes in the number with the threshold set to 0.95 and 0.80. As shown in the figure, the greatest alignment length obtained became long and the number of alignments obtained increased when the threshold value is small. Therefore, it is necessary to take the alignment length in the proposed score into consideration.

4. Conclusions and Future Work

We proposed a method for aligning metabolic pathways based on compound similarity. This method was applied to metabolic pathways in *E. coli*, and we found reaction similarities between fructose and mannose metabolic pathways and the galactose metabolic pathway. In future, it will be necessary to (1) develop a method for compensation of alignment score, (2) improve the similarity score between reactions, and (3) examine the pathway extraction algorithm.

Acknowledgments This study was supported in part by the "High-Tech Research Center" Project for Private Universities: matching fund subsidy from MEXT (Ministry of Education, Culture, Sports, Science, and Technology) 2005–2007, and the Ministry of Education, Science, Sports, and Culture, Grant-in-Aid for Young Scientists (B), 17700297, 2007.

References

1) Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M.: The KEGG resource for deciphering the genome, *Nucleic Acids Res.*, Vol.32, pp.D277–280 (2004).
2) Caspi, R., Foerster, H., Fulcher, C.A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee, S.Y., Tissier, C., Zhang, P. and Karp, P.D.: MetaCyc: A multiorganism database of metabolic pathways and enzymes, *Nucleic Acids Res.*, Vol.34, pp.D511–516 (2006).
3) Dandekar, T., Schuster, S., Snel, B., Huynen, M. and Bork, P.: Pathway alignment: Appliation to the comparative analysis of glycolytic enzymes, *Biochemical J.*, Vol.343, No.1, pp.115–124 (1999).
4) 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).
5) 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).
6) 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).
7) Webb, E.C. (Ed.): Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classifcation of Enzymes, Academic Press (1993).
8) 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).
9) 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).
10) MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.
11) Smith, T.F. and Waterman, M.S.: Identification of common molecular subsequences, *J. Mol. Biol.*, Vol.147, pp.195–197 (1981).
12) OpenEye Scientific Software: http://http://www.eyesopen.com/.
13) MacCuish, N.E. and MacCuish, J.D.: Clustering compound data: Asymmetric clustering of chemical datasets, chemometrics and cheminfomatics, *ACS Symposium Series*, Lavine, B.K. (Ed.), Vol.894, Oxford University Press, (2005).
http://www.mesaac.com/
14) Rahman, S.A., Advani, P., Schunk, R., Schrader, R. and Schomburg, D.: Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC), *Bioinformatics*, Vol.21, No.7, pp.1189–1193 (2005).

Appendix

List of MACCS 166 keys: In the list, "A" means any atom except "H," such as "C," "S," "O," and "N". "Q" means any atom except "H" and "C". "@" means that the atom of the number below @ is connected to the following atom. "%" indicates an aromatic bond. "$" means that the next bound is a ring bond. "!" means that the next bond is a chain bond.

Key	Description	Key	Description
1	ISOTOPE	84	NH2
2	103 < ATOMIC NO. < 256	85	CN(C)C
3	GROUP IVA,VA,VIA PERIODS 4-6 (GE..)	86	CH2QCH2
4	ACTINIDE	87	X!A$A
5	GROUP IIIB,IVB (SC..)	88	S
6	LANTHANIDE	89	OAAAO
7	GROUP VB,VIB,VIIB (V..)	90	QHAACH2A
8	QAAA@1	91	QHAAACH2A
9	GROUP VIII (FE..)	92	OC(N)C
10	GROUP IIA (ALKALINE EARTH)	93	QCH3
11	4M RING	94	QN
12	GROUP IB,IIB (CU..)	95	NAAO
13	ON(C)C	96	5M RING
14	S-S	97	NAAAO
15	OC(O)O	98	QAAAAA@1
16	QAA@1	99	C=C
17	CTC	100	ACH2N
18	GROUP IIIA (B..)	101	8M RING OR LARGER
19	7M RING	102	QO
20	Si	103	CL
21	C=C(Q)Q	104	QHACH2A
22	3M RING	105	A$A($A)$A
23	NC(O)O	106	QA(Q)Q
24	N-O	107	XA(A)A
25	NC(N)N	108	CH3AAACH2A
26	C$=C($A)$A	109	ACH2O
27	I	110	NCO
28	QCH2Q	111	NACH2A
29	P	112	AA(A)(A)A
30	CQ(C)(C)A	113	Onot%A%A
31	QX	114	CH3CH2A
32	CSN	115	CH3ACH2A
33	NS	116	CH3AACH2A
34	CH2=A	117	NAO
35	GROUP IA (ALKALI METAL)	118	ACH2CH2A>1
36	S HETEROCYCLE	119	N=A
37	NC(O)N	120	HETEROCYCLIC ATOM>1 (&..)
38	NC(C)N	121	N HETEROCYCLE
39	OS(O)O	122	AN(A)A
40	S-O	123	OCO
41	CTN	124	QQ
42	F	125	AROMATIC RING>1
43	QHAQH	126	A!O!A
44	OTHER	127	A$A!O>1 (&..)
45	C=CN	128	ACH2AAACH2A
46	BR	129	ACH2AACH2A
47	SAN	130	QQ>1 (&..)
48	OQ(O)O	131	QH>1
49	CHARGE	132	OACH2A
50	C=C(C)C	133	A$A!N
51	CSO	134	X (HALOGEN)
52	NN	135	Nnot%A%A
53	QHAAAQH	136	O=A.1
54	QHAAQH	137	HETEROCYCLE
55	OSO	138	QCH2A>1 (&..)
56	ON(O)C	139	OH
57	O HETEROCYCLE	140	O>43 (&..)
58	QSQ	141	CH3>2 (&..)
59	Snot%A%A	142	N>1
60	S=O	143	A$A!O
61	AS(A)A	144	Anot%A%Anot%A
62	A$A!A$A	145	6M RING>1
63	N=O	146	O>2
64	A$A!S	147	ACH2CH2A
65	C%N	148	AQ(A)A
66	CC(C)(C)A	149	CH3>1
67	QS	150	A!A$A!A
68	QHQH (&..)	151	NH
69	QQH	152	OC(C)C
70	QNQ	153	QCH2A
71	NO	154	C=O
72	OAAO	155	A!CH2!A
73	S=A	156	NA(A)A
74	CH3ACH3	157	C-O
75	A!N$A	158	C-N
76	C=C(A)A	159	O>1
77	NAN	160	CH3
78	C=N	161	N
79	NAAN	162	AROMATIC
80	NAAAN	163	6M RING
81	SA(A)A	164	O
82	ACH2QH	165	RING
83	QAAAA@1	166	FRAGMENTS

(Received April 20, 2007)
(Accepted July 9, 2007)

(Communicated by *Shigehiko Kanaya*)

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 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.