

A Multiple Alignment Algorithm for Metabolic Pathway Analysis using Enzyme Hierarchy

Yukako Tohsato Hideo Matsuda Akihiro Hashimoto

Department of Informatics and Mathematical Science
Graduate School of Engineering Science, Osaka University,
1-3 Machikaneyama, Toyonaka, Osaka 560-8531 Japan
E-mail: {yukako, matsuda, hasimoto}@ics.es.osaka-u.ac.jp
Tel: +81-6-6850-6601 Fax: +81-6-6850-6602

Keywords: alignment, metabolic pathway, pathway analysis, enzyme, EC number

Abstract

In many of the chemical reactions in living cells, enzymes act as catalysts in the conversion of certain compounds (substrates) into other compounds (products). Comparative analyses of the metabolic pathways formed by such reactions give important information on their evolution and on pharmacological targets (Dandekar *et al.* 1999). Each of the enzymes that constitute a pathway is classified according to the EC (Enzyme Commission) numbering system, which consists of four sets of numbers that categorize the type of the chemical reaction catalyzed. In this study, we consider that reaction similarities can be expressed by the similarities between EC numbers of the respective enzymes. Therefore, in order to find a common pattern among pathways, it is desirable to be able to use the functional hierarchy of EC numbers to express the reaction similarities. In this paper, we propose a multiple alignment algorithm utilizing information content that is extended to symbols having a hierarchical structure. The effectiveness of our method is demonstrated by applying the method to pathway analyses of sugar, DNA and amino acid metabolisms.

Introduction

In many of the chemical reactions in living cells, enzymes act as catalysts in the conversion of certain compounds (substrates) into other compounds (products). In recent years, these chemical reactions have been the subjects of much research, and reaction data are being accumulated in databases such as WIT (Overbeek *et al.* 2000) and KEGG (Ogata *et al.* 1999), which are open to the public via the World Wide Web. Generally, a series of such chemical reactions is called a pathway. Pathways may have many branches even if they are fundamental pathways such as the sugar metabolism (see Fig. 1).

Some examples of the applications of computer analysis to pathways are (1) pathway reconstruction (Mavrouniotis 1993; Gaasterland and Selkov 1995;

Goto *et al.* 1997) (where the input consists of substrates and products for each of enzymes, and the output is a set of potential pathways); (2) pathway clustering (Forst and Schulten 1999) (where the distances between pathway pairs are calculated by aligning enzymes, and pathways are classified based on the distances); and (3) genome comparison based on pathway data (Dandekar *et al.* 1999; Bono *et al.* 1998) (where genomes are compared by assigning genes on the genomes to each of enzymes which constitutes a specific pathway).

In this study, we consider the multiple alignment of more than two pathways as one of the techniques for the comparative analysis of metabolic pathways based on reaction similarity. The comparison of pathways is usually performed based on the sequence similarity between enzymes, such as in pathway clustering. However, it has been found that comparison based on sequence similarity is not necessarily appropriate, because reaction similarity does not necessarily correlate with sequence similarity due to the enzyme recruitment (Galperin, Walker, and Koonin 1998). Therefore, in this study, we intend to perform the multiple alignment of pathways based on the enzyme hierarchy.

Enzyme hierarchy and reaction similarities

Each of the enzymes that constitute a pathway is characterized by the reactions that they catalyze. The International Union of Biochemistry and Molecular Biology (IUBMB) has developed a classification scheme based on this observation. The scheme is hierarchical, with four levels. At the top of the hierarchy are six broad classes of enzymatic activity (indicated by the first number of the EC number) (Enzyme Nomenclature 1992).

1. Oxidoreductase
2. Transferase

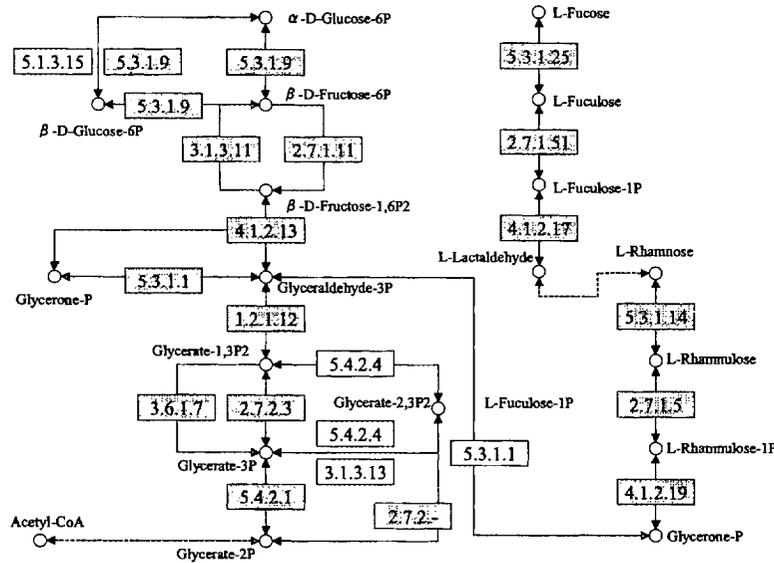


Figure 1: Sugar degradation pathway. The enzymes are shown in boxes with the EC numbers inside. The shaded boxes represent those enzymes whose genes are identified in *Escherichia coli*. Pathways indicated by dotted lines are not shown. Acetyl-CoA is the final product in the pathway.

3. Hydrolase
4. Lyase
5. Isomerase
6. Ligase

EC numbers are generally expressed as a string of four numbers separated by periods.

For example, consider EC number identified as 1.1.1.1. This class is a member of the top level group [1], the oxidoreductases. The second level of the hierarchy identifies a subclass; for the oxidoreductases, the second level specifies the kind of donor which is oxidized. In this case, sub-class [1.1] means the enzyme acts on the CH-OH group of donors. The third position in the oxidoreductase group specifies the kind of acceptor. In this example, the sub-sub-class [1.1.1] means that NAD⁺ or NADP⁺ are the acceptor. The lowest level in the hierarchy (the leaf node [1.1.1.1]) identifies a particular reaction. Therefore the donor for [1.1.1.1] and [1.1.1.2] is CH-OH and their acceptors are NAD⁺ and NADP⁺, respectively.

We consider five class levels, including a new top class [*] which we have added to express arbitrary enzymes (see Fig. 2). References to EC numbers with fewer than four numbers indicate an internal node in the tree, including all of the subclasses or leaves below it. The numbers specify a path down the hierarchy,

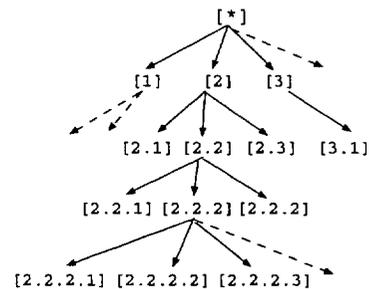


Figure 2: Example of enzyme hierarchy.

with the leftmost number identifying the highest level. In this paper, enzymes are denoted by their enzyme classes (e.g. [2.2.2.3]).

Fig. 3 shows examples of reaction similarities. For each of enzymes which catalyze similar reactions in these pathways, similar pairs of EC numbers occurs ([2.4.2.3] - [2.4.2.4], [3.1.3.5] - [3.1.3.5] and [2.7.4.14] - [2.7.4.9]), and their positions within the enzyme hierarchy are near. Therefore, in this study, we consider that the relationship between the proximity within the enzyme hierarchy and the reaction similarity is strong, and we express the reaction similarities between pathways based on enzyme hierarchy.

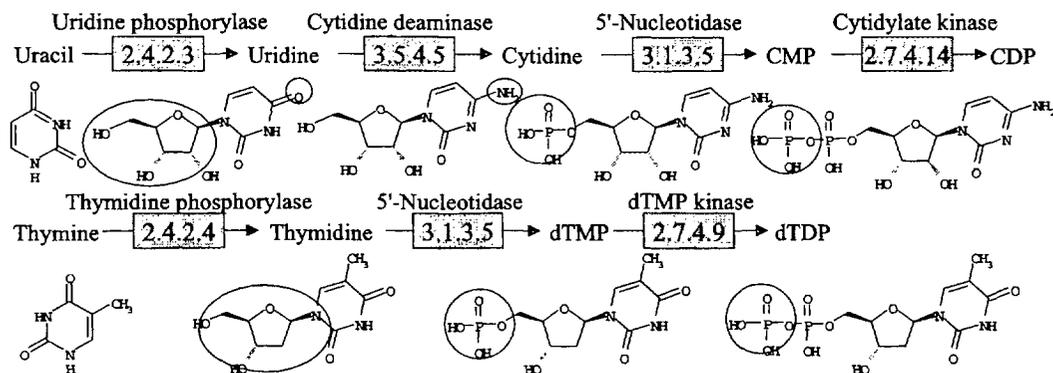


Figure 3: Structural comparison of similar pathways. The enzymes' name are shown above their EC numbers, and their substrates and products are shown between the enzymes. The structural components that are modified by the enzymes are circled.

An Information Content Maximization Alignment Algorithm

Similarity score between enzymes

Although reaction similarity is closely related to the proximity of the enzymes within the enzyme hierarchy, there is a problem with expressing the enzyme similarity only by the proximity within the enzyme hierarchy (for example, by considering the numeric similarity between the two enzyme numbers). This is because there is a large deviation in the distribution among the enzyme hierarchy. For example, the EC numbers located under [1.1.1] range from 1.1.1.1 to 1.1.1.254, whereas there is only one EC number 5.3.4.1 located under [5.3.4].

Moreover, when the enzymes which occupy a near position in the enzyme hierarchy are included in a set of pathways to align, these enzymes will be judged with similar group of the enzymes, even if these enzymes are combined at random.

In consideration of above points, in this study, we decide the degree of similarity between the enzymes as follows. Some definitions required to express this degree of similarity are shown below. Unless specified otherwise, we represent enzymes solely by EC numbers in the definitions.

Definition 1 The hierarchy constructed by using the EC numbering system (described in the preceding section) is called the **enzyme hierarchy**. Also each element of the enzyme hierarchy (e.g. [2.2.2.1], [2.2.1] and [*]) is called the **enzyme class**.

Definition 2 Given more than two enzymes as input, the enzyme class which is the lowest class in the upper classes of those enzymes on the enzyme hierarchy is

called **common upper class**. Between the same enzymes, their common upper class is the same as their enzyme class.

For example, [1.2.3], [1.2] and [*] are common upper classes between [1.2.3.4] and [1.2.3.5], [1.2.3.4] and [1.2.4.1], and [1.2.3.4] and [2.1.1.1], respectively.

Definition 3 For an enzyme class h , we denote by $E(h)$ the set of all enzymes whose classes are included under the enzyme class h . $C(h)$ expresses the number of elements of $E(h)$.

Definition 4 For a set of pathways $S = \{s_1, \dots, s_n\}$, $N(s_i)$ is the number of enzymes which appear in element s_i ($1 \leq i \leq n$) of S , and $o(e, s_i)$ is the number of times that the enzyme e appears in s_i . Then

$$p(e) = \sum_{i=1}^n o(e, s_i) / \sum_{i=1}^n N(s_i) \quad (1)$$

is called the **occurrence probability** of e in S .

Definition 5 Given a set of pathways S , for the enzyme class h , $p(h)$ is called the **occurrence probability** of h in S , and represents the sum of the occurrence probabilities of all enzymes in $E(h)$.

Definition 6 For an enzyme class h , $I(h)$ is expressed with the following formula is called the **information content** of h .

$$I(h) = \log_2 \frac{1}{C(h)} - \log_2 p(h) \quad (2)$$

For two enzymes e_i, e_j , if their common upper class is h_{ij} , then we consider $I(h_{ij})$ express the similarity between e_i and e_j (the value increases with greater similarity).

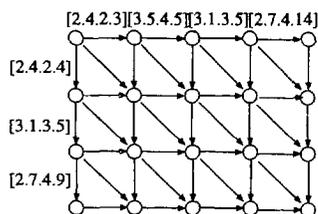


Figure 4: The global dynamic program matrix for two pathways, with arrow indicating paths of alignment between two pathways.

Although the above value of the information content depends on the given pathway set, the information content of the highest rank $I([\ast])$ is a fixed constant $\log_2(1/3705) - \log_2(1) = -11.85$ since there are 3705 kinds of EC number according to the KEGG database and $p([\ast]) = 1$. Moreover, for any enzyme class h , $I(h) \geq I(h_i)$ ($1 \leq i \leq n$) holds where h_1, h_2, \dots are sub-classes of h .

The information content defined here corresponds with the score matrix used by the dynamic programming alignment algorithm as described below.

The degree of similarity between two any amino acids a and b can be expressed by the following formula (Durbin *et al.* 1998).

$$S(a, b) = \log \frac{p_{ab}}{q_a q_b} = \log(p_{ab}) - \log(q_a q_b) \quad (3)$$

When a common ancestor exists between a and b (it may be the same as a or b), the numerator p_{ab} is the probability that a is replaced by b (namely, the probability that a and b relate evolutionarily), and the denominator expresses the product of the probabilities of obtaining a and b , respectively, by substitution at random (namely, the probability with which a and b are produced independently). That is, this score expresses the degree to which a and b relate evolutionarily in terms of a log-odds ratio.

We consider that the information content which is defined in this study is the log-odds ratio between the probability with which two enzymes are similar in the enzyme hierarchy (i.e., the number of all enzymes under their common upper class is small) and the probability with which two enzymes match at random in a set of pathways.

Pairwise alignment

In this study, to create a pattern from the optimal pairwise alignment with hierarchical structure, we extend the global alignment algorithm based on dynamic programming (Needleman and Wunsch 1970). The

	[2.4.2.3]	[3.5.4.5]	[3.1.3.5]	[2.7.4.14]
[2.4.2.4]	[2.4.2]	[*]	[*]	[2]
[3.1.3.5]	[*]	[3]	[3.1.3.5]	[*]
[2.7.4.9]	[2]	[*]	[*]	[2.7.4]

Figure 5: Common upper class reference table for two pathways [2.4.2.3] [3.5.4.5] [3.1.3.5] [2.7.4.14] and [2.4.2.4] [3.1.3.5] [2.7.4.9].

Pathway1	Pathway2	Pattern	MatchScore
[2.4.2.3]	[2.4.2.4]	[2.4.2]	-3.40
[3, 5, 4, 5]		"_"	-15
[3.1.3.5]	[3.1.3.5]	[3.1.3.5]	1.80
[2.7.4.14]	[2.7.4.9]	[2.7.4]	-2.51

Figure 6: Example of the optimal alignment between two pathways, showing construction of a covering pattern from the common upper class covering both aligned elements (see Fig. 5) and the resulting information content between aligned elements.

global alignment based on dynamic programming arranges the elements in each pair of sequences in two dimensions, and looks for the optimal path which starts at the upper-left point, and reaches the bottom-right point, which is the goal point.

The alignment score is calculated as follows. When a diagonal arrow is selected, the information content of the common upper class (see Fig. 5) that corresponds to the arrow is added. When a left-to-right or top-to-bottom arrow is selected, the information content of a gap is added. Here, the optimal alignment is the alignment which takes the highest score. A gap in a pathway expresses any enzyme. We consider that it is desirable that the information content of a gap is less than the information content of $[\ast]$, and thus we set the information content of a gap to -15 .

We define that the sequence of the common upper classes and gaps that correspond to the arrows in the optimal alignment is the **pattern** corresponding to the alignment. The information content $I(p)$ of pattern p is the score of the alignment.

For example, given two pathways [2.4.2.3] [3.5.4.5] [3.1.3.5] [2.7.4.14] and [2.4.2.4] [3.1.3.5] [2.7.4.9], the pattern for the optimal alignment is [2.4.2] "_" [3.1.3.5] [2.7.4], as shown in Fig. 6. In this case, the gap "_" signifies one insertion into Pathway2. Thus, the information content of the pattern [2.4.2] "_" [3.1.3.5] [2.7.4] in Fig. 6 has a total value of $-3.40 - 15 + 1.80 - 2.51 =$

This procedure's complexity is $O(l^2)$, where l is the maximum length of two pathways.

Extension to multiple alignment

We consider the extension of the pairwise alignment algorithm to allow alignment of more than two pathways. To obtain multiple alignment, we introduce a pairwise alignment algorithm for two patterns obtained by the pairwise alignment for two pathways. The result of the algorithm is also a pattern. Given a set of more than two pathways, the multiple alignment performs a pairwise alignment between any two elements in a set of patterns, replaces the two patterns by the pattern obtained from the pairwise alignment, and repeats the pairwise alignment until the optimal multiple alignment is obtained. For each pairwise alignment, we are assured that the size of the set decreases by one.

To find the optimal multiple alignment, we maximize the information content of the set of patterns. However, it is not desirable to simply sum the information content of all patterns in the set. This is because the information content decreases as the number of patterns decreases due to repetition of the pairwise alignment. Moreover for each pairwise alignment, the enzyme classes in the obtained pattern tend to be one level higher in the enzyme hierarchy compared to these in the original pair. To solve these problems, we consider that the difference in information content between enzyme classes that differ one level is w , and thus we add w to the information content of the pattern for each execution of the pairwise alignment.

Specifically, given a set of pathways S , the information content $I(P)$ of an output set P of patterns will be calculated by following formula:

$$I(P) = w\bar{l}(n - k) + \sum_{p \in P} \frac{n_s}{n} I(p), \quad (4)$$

where $n = \#S$, $k = \#P$ (for a set X , we denote by $\#X$ the number of elements of X), n_s is the number of patterns in S , p is a pattern obtained from the multiple alignment, the constant \bar{l} is the average length of the given set of pathways, and constant w is the weight for an element of alignment. $w\bar{l}(n - k)$ is determined so that the number of elements in a set of patterns P is small. We found that the information content differed by only 5 for enzyme classes which differed by one class. So, in this paper, we set $w = 5$.

If we attempt to find the multiple alignment having the maximum score among all possible alignments, we need a multi-dimensional alignment, and the complex-

procedure IMA

```

input: A set of pathways  $S$ ;
output: A set of patterns  $P$ ;
 $n := \#S$ ;  $P := S$ ;  $P_0 := S$ ;  $max := I(P)$ ;
for  $k := n - 1$  downto 1 do
  foreach  $p_i \in P$  do
    foreach  $p_j \in P$  ( $p_i \neq p_j$ ) do
       $p := G(p_i, p_j)$ ;
       $P' := (P - \{p_i, p_j\}) \cup \{p\}$ ;
      if  $max < I(P')$  then
         $P_{n-k} := P'$ ;
         $max := I(P')$ ;
      endforeach
    endforeach
  if  $I(P_{n-k}) < I(P_{n-k-1})$  then
    return  $P$ ;
   $P := P_{n-k}$ ;
endfor
return  $P$ ;

```

Figure 7: Multiple alignment algorithm for pathways: $p := G(p_i, p_j)$ is a procedure which performs pairwise alignment of two pathways p_i and p_j and calculates a resulting pattern p .

ity becomes

$$O\left(\sum_{k=1}^{n-1} \binom{n}{n-k+1} l^{n-k+1}\right).$$

This complexity is not practical for large n . The multiple alignment for sequences usually uses a greedy algorithm (Feng and Doolittle 1987). Therefore, we implement the multiple alignment algorithm based on a greedy algorithm (see Fig. 7).

Given n pathways, the procedure which obtains $k = n - 1$ patterns is the operation by which any two patterns is chosen from those sets, the alignment is executed for those patterns, and the pattern with the largest information content is selected.

The next procedure which obtains $k = n - 2$ patterns takes the previously-obtained result (whose set is the original with two patterns removed and one added) and executes the same operation. The procedure stops when $k = 1$ (as a result, one solution pattern is obtained). If the process repeats; i.e., if these procedures cannot obtain a set of patterns whose information content is larger than the information content of the previous set of patterns (P_{n-k-1} in Fig. 7), it returns the previous set of patterns as a solution, and the procedure stop.

This procedure's complexity is $O(n^3 l_{max}^2)$, where l_{max} is the maximum length of pathway in the set of

5.3.1.9	2.7.1.11	4.1.2.13
5.3.1.25	2.7.1.51	4.1.2.17
5.3.1.14	2.7.1.5	4.1.2.19
5.3.1.8	2.7.1.11	4.1.2.13

Figure 8: Glucose, fucose, rhamnose and mannose degradation pathways in *Escherichia coli*.

pathways S as input, and $n = \#S$.

Experiments

The effectiveness of our method is demonstrated by applying this algorithm to the sugar, DNA and amino acid metabolic pathways extracted from the metabolism map of KEGG. Since the algorithm shown in Fig. 7 does not consider branching pathways that occur in those metabolic pathways, we need a pre-processing procedure which extracts a set of non-branching pathways that exhibit reaction similarities as shown in Fig. 3. In this paper, we performed the pre-processing procedure as follows.

First, we search for all possible pairs of non-branching pathways whose length is three or more, which show reaction similarities to each other. Since the number of the possible pairs could become too large to handle in our algorithm, we reduce the number by limiting alternative branches of pathways to the most suitable one. Here, by 'suitable', we mean the branch of pathways which includes ligands having a larger number of carbons than those included in the other branches.

Secondly, we combine the pairs of pathways into a set of pathways by linking them with common pathways (i.e., a pair of pathway A and pathway B, and a pair of pathway B and C are linked together to a set of pathways A, B and C).

Finally, we inspect the resulting set to select the input set of the algorithm shown in Fig. 7. In our current implementation, this step is not fully automated but partially performed by eye inspection.

The algorithm was coded and the experiments were performed using a DEC AlphaStation 600 5/333 (Alpha 21164 333MHz, Tru64 UNIX, gcc).

The extracted pathways are shown using matrices in Fig. 8, Fig. 9 and Fig. 10. In Fig. 8, each row in a matrix expresses one pathway as a sequence of EC numbers representing its enzymes.

In Fig. 8, the pathways starts from enzymes which take α -D-Glucose-6P, L-Fucose, L-Rhamnose and D-Mannose-6P as initial substrates, respectively. Each of the enzymes corresponds with ORF(Open Reading Frame) in *Escherichia coli* according to KEGG.

For this data, we obtained one pattern [5.3.1] [2.7.1]

Purine base	2.4.2.1	3.1.3.5	2.7.4.3	2.7.4.6	2.7.7.7	DNA	
	2.4.2.1	3.1.3.5	2.7.4.3	2.7.1.40	2.7.7.7		
	2.4.2.1	3.1.3.5	2.7.4.8	2.7.4.6	2.7.7.7		
	2.4.2.1	3.1.3.5	2.7.4.8	2.7.1.40	2.7.7.7		
	2.4.2.1	3.1.3.5	2.7.4.3	2.7.4.6	2.7.7.6		
Pyrimidine base	2.4.2.1	3.1.3.5	2.7.4.3	2.7.1.40	2.7.7.6	RNA	
	2.4.2.1	3.1.3.5	2.7.4.8	2.7.4.6	2.7.7.6		
	2.4.2.1	3.1.3.5	2.7.4.8	2.7.1.40	2.7.7.6		
	2.4.2.4	3.1.3.5	2.7.4.9	2.7.4.6	2.7.7.7		DNA
	2.4.2.1	3.5.4.5	3.1.3.5	2.7.4.14	2.7.4.6		
Pyrimidine base	2.4.2.3	3.5.4.5	3.1.3.5	2.7.4.14	2.7.4.6	2.7.7.6	RNA
	2.4.2.4	3.1.3.5	2.7.4.14	2.7.4.6	2.7.7.6		

Figure 9: DNA and RNA replication pathways in *Es-*

Eco	+ 4.2.1.17	1.1.1.35	2.3.1.16			
	+ 4.2.1.17	1.1.1.57	2.3.1.9			
	+ 4.2.1.17	1.1.1.35	2.3.1.9			
Afu	+ 1.3.99.3	4.2.1.17	1.1.1.35	2.3.1.16		
	1.2.4.2	2.3.1.61	1.3.99.7	4.2.1.17		1.1.1.35
Cel	1.2.4.2	1.3.99.7	4.2.1.17	1.1.1.35	2.3.1.9	
	+ 2.3.1.2	1.3.99.3	4.2.1.17	1.1.1.35		
	+ 2.3.1.2	1.3.99.6	4.2.1.17	1.1.1.35		
	+ 1.3.99.2	4.2.1.17	1.1.1.35			

Figure 10: Isoleucine, lysine, tryptophan and other degradation pathways in different organisms: Eco, *Escherichia coli*; Afu, *Archaeoglobus fulgidus*; Cel, *Caenorhabditis elegans*. The pathways marked by "+" do not have a complete pathway from the initial substrate.

[4.1.2] by applying the algorithm in Fig. 7. As described before, we consider that the hierarchy of EC number of enzyme expresses the similarity of a reaction.

This result is adequate, because the enzyme classes in the enzyme hierarchy are upper than those in Fig. 8. In addition, the information content of the obtained solution is 32.56, and the calculation time of the alignment is 0.005 seconds.

In Fig. 9, the pathways start from enzymes which take purine or pyrimidine as substrates, replicate DNA or RNA (the last 2.7.7.7 and 2.7.7.6 are polymerases for DNA and RNA, respectively). Each of the enzymes corresponds with ORF in *Escherichia coli* according to KEGG.

For this data, we obtained one pattern [2.4.2] [3.1.3.5] [2.7.4] [2.7] [2.7.7] by applying the algorithm in Fig. 7.

Just before this pattern is finally obtained by the for-loop of the algorithm in Fig. 7, the alignment of the patterns [2.4.2] [3.1.3.5] [2.7.4] [2.7] [2.7.7] and [2.4.2] [3.5.4.5] [3.1.3.5] [2.7.4.14] [2.7.4.6] [2.7.7] is performed, and the information content increases from 239.67 to 245.94 at this time. This result signifies that a part of DNA and RNA replication pathways includes the reac-

tion catalyzed by [3.5.4.5] by the alignment of pathway including gap. In addition, the information content of the set of obtained patterns is 245.94, and the calculation time of the alignment is 0.15 seconds.

In Fig. 10, each of the pathways are parts of pathways which degrade the amino acids such as isoleucine, lysine, tryptophan and other degradation pathways. Each of the enzymes corresponds with ORF in *Escherichia coli*, *Archaeoglobus fulgidus* and *Caenorhabditis elegans* according to KEGG.

For this data, we obtained three patterns

```
{ [4.2.1.17] [1.1.1.] [2.3.1],  
  "-" [1.3.99] [4.2.1.17] [1.1.1.35] "-",  
  [1.2.4.2] "-" [1.3.99.7] [4.2.1.17] [1.1.1.35] [2.3.1.9]  
}
```

by applying the algorithm in Fig. 7.

Among these patterns, the first pattern corresponds to the three pathways in *Escherichia coli*.

Although the algorithm of Fig. 7 would obtain two patterns "-" "-" [1.3.99] [4.2.1.17] [1.1.1.35] "-" and [4.1.2.17] "-" [1.1.1] [2.3.1] in the execution process combining these three patterns, the information content would decrease from 56.70 to 52.46, and thus the repetition is finished in this stage. Therefore, the information content of the set of obtained patterns is 56.70, the calculation time of the alignment is 0.04 seconds.

Conclusion

We propose a multiple alignment algorithm for pathways utilizing information content. The effectiveness of our method is demonstrated by applying the method to pathway analyses of sugar, DNA and amino acid metabolisms. In this experiment, we collected experimental data on pathways by eye-inspection. In future, this portion needs to be automated. Moreover, in the application of our method to pathway analysis, we are working to check the validity of information content as a measure of reaction similarity.

Acknowledgments

This work was supported in part by ACT-JST and CREST of JST (Japan Science and Technology), and a Grant-in-Aid "Genome Science" (08283103) for Scientific Research on Priority Areas from the Ministry of Education, Science, Sports and Culture in Japan.

References

- Overbeek, R., Larsen, N., Pusch, G. D., D'Souza, M., Selkov, E. Jr, Kyrpides, N., Fonstein, M., Maltsev, N., and Selkov, E. 2000. WIT: Integrated System for High-Throughput Genome Sequence Analysis and Metabolic Reconstruction, *Nucleic Acids Research*, 28(1):123–125 (Available at: <http://wit.mcs.anl.gov/WIT2/>).
- Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Research*, 27(1):29–34 (Available at: <http://www.genome.ad.jp/kegg/>).
- Dandekar, T., Schuster, S., Snel, B., Huynen, M. and Bork, P. 1999. Pathway Alignment: Application to the Comparative Analysis of Glycolytic Enzymes, *Biochemical J.*, 343(1):115–124.
- Mavrovouniotis, M. L. 1993. Identification of Qualitatively Feasible Metabolic Pathways, *Artificial Intelligence and Molecular Biology* (Hunter, L. (ed.)), AAI Press/MIT Press, Menlo Park, 325–364.
- Gaasterland, T. and Selkov, E. 1995. Reconstruction of Metabolic Networks Using Incomplete Information, *Proc. Intl. Conf. on Intelligent Systems for Molecular Biology*, 127–135.
- Goto, S., Bono, H., Ogata, H., Fujibuchi, W., Nishioka, T., Sato, K. and Kanehisa, M. 1997. Organizing and Computing Metabolic Pathway Data in terms of Binary Relations, *Pacific Symp. Biocomputing 97*, 175–186.
- Bono, H., Ogata, H., Goto, S. and Kanehisa, M. 1998. Reconstruction of Amino Acid Biosynthesis Pathways from the Complete Genome Sequence. *Genome Research*, 8(3):203–210.
- Forst, V. C. and Schulten K. 1999. Evolution of Metabolisms: A New Method for the Comparison of Metabolic Pathways Using Genomics Information, *J. Computational Biology*, 6(3):343–360.
- Galperin, M. Y., Walker, D. R. and Koonin, E. V. 1998. Analogous Enzymes: Independent Inventions in Enzyme Evolution, *Genome Research*, 8(8):779–790.
- Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes, NC-IUBMB. Academic Press, New York, NY.
- Durbin, R., Eddy, S. R., Krogh, A. and Mitchison, G. 1998. *Biological Sequence Analyses: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge.
- Needleman, S. B. and Wunsch, C. D. 1970. A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins, *J. Mol. Biol.*, 48:444–453.

Feng, D. and Doolittle, R. F. 1987. Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees, *J. Mol. Evol.*, 25:351–360.