A Local Alignment Algorithm for Metabolic Pathway Analysis

Yukako Tohsato
yukako@ics.es.osaka-u.ac.jp

Hideo Matsuda
matsuda@ics.es.osaka-u.ac.jp

Akihiro Hashimoto
hasimoto@ics.es.osaka-u.ac.jp

Department of Informatics and Mathematical Science,
Graduate School of Engineering Science, Osaka University,
1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan

Keywords: alignment, metabolic pathway, enzyme, pathway analysis

1 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). Comparative analyses of the metabolic pathways formed by such reactions give important information on their evolution and on pharmacological targets [1]. 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 (see Fig. 1). 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 (local) 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.

2 Method

Given two pathways like Fig. 1, we consider finding the local alignment using the enzyme tree structure such as Fig. 2. We extend the local alignment algorithm[2] based on dynamic programming (see Fig. 3). Matrix N in Fig. 3 stores the nearest common ancestor between the enzymes. [2.4.2] and [2.7.4] are common ancestor enzyme classes between [2.4.2.4] and [2.4.2.3], and between [2.7.4.9] and [2.7.4.14], respectively. Graph M in Fig. 3 expresses the combination of sub alignments and traces the optimum alignment between two pathways. The result of the alignment is constituted by enzymes (e.g. [3.1.3.5]) and enzyme classes (e.g. [2.4.2]), such as "[3.1.3.5] [2.7.4]". The recurrence equation for the algorithm is:

$$M_{i,j} = \max\left\{0, M_{i-1,j-1} + I(\{h_{1i}, h_{2j}\}), M_{i-1,j}, M_{i,j-1}\right\}. \quad (1)$$

Given the nearest common ancestor h between enzyme h_{1i} and h_{2j}, the information content $I(\{h_{1i}, h_{2j}\})$ $= I(h)$ is the logarithm of the occurrence probability of enzyme class h in the pathways. The information content I([*]) of the top enzyme class [*] is always 0. $I(p)$ is the sum of the information contents of enzymes and enzyme classes composing an optimum alignment p.

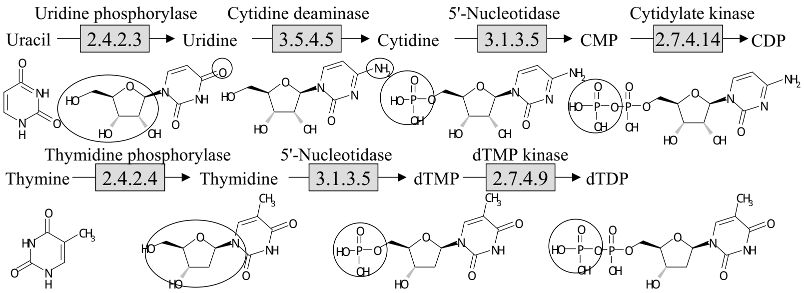

Figure 1: Structural comparison of similar pathways.

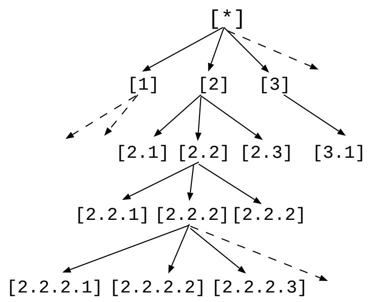

Figure 2: Example of enzyme hierarchy.

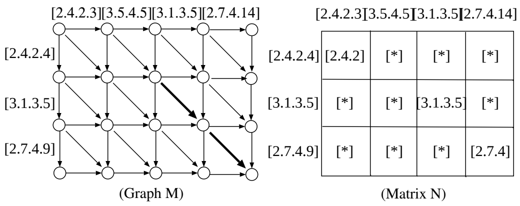

Figure 3: Our extended local alignment algorithm.

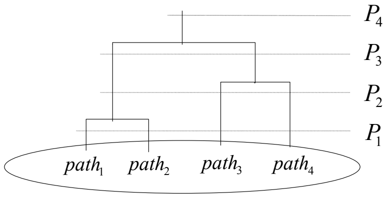

Figure 4: Multiple alignment procedure.

When more than two sequences are given as input, we use a greedy algorithm for aligning those sequences (see Fig. 4). Then, we introduce an evaluation function to express a family of pathways as follows:

$$I(P) = (-\log_2 \frac{k}{n}) \cdot (\sum_{p \in P} \frac{n_s}{n} I(p)) \tag{2}$$

3 Application of Sugar, DNA and Amino Acid Metabolisms

We apply this algorithm to the sugar, DNA and amino acid metabolic pathways extracted from the metabolism map of KEGG[3]. The extracted pathways are shown in Fig. 5, Fig. 6, and Fig. 7. In these figures, matrices express the pathways. Each row of a matrix expresses one pathway as a sequence of EC numbers representing enzymes. In our experiment, we obtained an alignment "[5.3.1] [2.7.1] [4.1.2]" for Fig. 5, two alignments "[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]" for Fig. 6, and two alignments "[4.2.1.17] [1.1.1] [2.3.1]" and "[1.3.99] [4.2.1.17] [1.1.1.35]" for Fig. 7.

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

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		RNA
	2.4.2.1	3.1.3.5	2.7.4.3	2.7.1.40	2.7.7.6		
	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		
Pyrimidine base	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	2.7.7.7	
	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 6: DNA and RNA replication pathways in *Escherichia coli*

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		
Cel		1.2.4.2	2.3.1.61	1.3.99.7	4.2.1.17	1.1.1.35	2.3.1.9
		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 7: Isoleucine, lysine, tryptophan and other degradation pathways in different organisms.

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.*, 343(1):115–124, 1999.

[2] Smith, T.F. and Waterman, M.S., Identification of common molecular subsequences, *J. Mol. Biol.*, 147(1):195–197, 1981.

[3] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M., KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Research*, 27(1):29–34, 1999 (Available at: http://www.genome.ad.jp/kegg/).