Original Paper

A Method for Species Comparison of Metabolic Networks Using Reaction Profile

Yukako Tohsato†

Comparative analyses of the metabolic networks among different species provide important information regarding the evolution of organisms as well as pharmacological targets. In this paper, a method is proposed for comparing metabolic networks based on enzymatic reactions within different species. Specifically, metabolic networks are handled as sets of enzymatic reactions. Based on the presence or absence of metabolic reactions, the metabolic network of an organism is represented by a bit string comprised of the digits "1" and "0," called the "reaction profile." Then, the degree of similarity between bit strings is defined, followed by clustering of metabolic networks by different species. By applying our method to the metabolic networks of 33 representative organisms selected from bacteria, archaea, and eukaryotes in the MetaCyc database, a phylogenetic tree was reconstructed that represents the similarity of metabolic network based on metabolic phenotypes.

1. Introduction

To obtain the energy necessary for cellular activities, cells within organisms take up many kinds of material in the form of food, etc. 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 the action of enzymes (proteins). As the product of a reaction is used as the substrate of other reactions, a large-scale and complex metabolic network is formed. These reactions are now stored in several public databases, including KEGG [1] and MetaCyc [2], which are available on the World Wide Web. For example, KEGG is a collection of manually drawn metabolic maps.

Metabolisms are important targets for understanding cell processes. Comparative analysis of the metabolic networks among different species provides essential information on the evolution of organisms and on pharmacological targets, and there has been a great deal of research in this area in recent years. Some examples of the application of computer analysis to metabolic networks include (1) metabolic pathway clustering based on genomic information [3),4)], where metabolic pathways are compared by assigning genes on the genomes to each of the enzymes that constitute a specific pathway and (2) pathway alignment based on functional similarity among enzymes [5] where metabolic pathways are clustered by assigning enzymes based on an enzymatic hierarchy. On the other hand, phylogenetic classification based on single genes, such as rRNA [6], does not provide a complete and accurate picture of evolution because it does not take into account evolutionary leaps due to gene transfer, duplication, deletion, and functional replacement [7]. Thus, importance is placed on comparison and investigation of phylogenetic trees created from a variety of standpoints [8]. Here, I focus on comparison of phylogenetic trees between metabolic phenotypes and genomic sequences, and a variety of related proposals have been made [8)~10)].

In this study, a method was developed for comparing different species based on metabolic expression profiles. The method consists of comparing different species to acquire new knowledge regarding interspecies phylogenies. This is done by considering the metabolic network as a set of metabolic reactions, and the whole network is expressed as a bit string encoding the presence (1) and absence (0) of reactions that comprise the network.

Section 2 discusses related research. The proposed method is described in Section 3, and Section 4 presents results obtained using the proposed method in an actual comparison of metabolic networks. Finally, Section 5 summa-

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

rizes problems and future work.

2. Related Work

There have been previous studies related to the purpose of this research, as reported by Hong, et al. [8]. They proposed a method for species comparison of metabolic networks based on the combination of metabolites that comprised the enzymatic reaction, and analyzed the metabolic pathways of 42 microorganisms. The method proposed by Hong, et al. classifies the overall metabolic networks into 64 individual sub-networks on the basis of metabolic map classifications. The numbers of reactions involved in each sub-network were counted and used for estimation of the reaction content p_{ij}.

$$p_{ij} = 100 \times r_{ij}/R_j \quad (1)$$

where r_{ij} is the number of reactions in the jth sub-network in organism i, and R_j is the number of non-duplicate reactions involved in the jth sub-network. In this case, the reaction content becomes synonymous with a sub-network. The Pearson correlation coefficient was used to assess the degree of similarity D between the reaction content $p_{i1}, p_{i2}, \cdots, p_{iN}$ of organism i with the reaction content $p_{j1}, p_{j2}, \cdots, p_{jN}$, which is defined as:

$$D = \frac{1}{N} \sum_{k=1,N} (\frac{p_{ik} - \bar{p_i}}{\sigma_i})(\frac{p_{jk} - \bar{p_j}}{\sigma_j}) \quad (2)$$

where $\bar{p_i}$ and $\bar{p_j}$ are the averages of values in $p_{i1}, p_{i2}, \cdots, p_{iN}$ and $p_{j1}, p_{j2}, \cdots, p_{jN}$, respectively. σ_i and σ_j are the standard deviations of these values. Then, clustering was performed using the furthest neighbor method [8].

Nevertheless, with this definition of Hong's method, if the numbers of reactions in the metabolic map are identical, these will not be distinguished, even where the types of reaction are different. Consider the case where, within organisms S_1, S_2, and S_3, the presence or absence of enzymatic reactions r_1, r_2, r_3, and r_4 have relationships shown in **Fig. 1**. The reaction contents of two organisms S_2 and S_3 become the score value 75; even where the existing enzyme reactions are of different types, they thus become the same score. In addition, the clustering results from the method of Hong, et al. are impacted by the metabolic map classifications.

	reactions			
	r_1	r_2	r_3	r_4
organism S_1	1	1	1	1
organism S_2	1	1	1	0
organism S_3	0	1	1	1

Fig. 1 The reaction profile of three organisms in the string that encodes reaction's presence or absence in an organism.

Taking the points described above into consideration, a method is proposed as described in the following section.

3. Method

3.1 Metabolic Network and Reaction Profile

The metabolic network of an organism is treated as a set of the enzymatic reactions. Consider two different organisms, S and S', with metabolic networks N and N', respectively. The set R of all reactions included within the metabolic networks of organisms S and S' is taken as $R = \{r_1, r_2, \cdots, r_n\}$. Here, multiple isozymes catalyzing the same reaction were counted only once, and multifunctional enzymes were counted as many times as they catalyze different reactions. Enzymatic reactions are distinguished by the combinations of metabolites —— called "reaction types." Duplication of reaction types is not allowed within a set R. For R, the reaction profile of organism X is represented by a bit string $Px = b_{x1}b_{x2}\cdots b_{xn}$ (a sequence of digits "0" and "1"). When a bit b_{xi} is set to 1 in a bit string, it means the corresponding reaction r_i $(1 \leq i \leq n)$ is present for organism X, while 0 means the reaction is absent.

3.2 Similarity Measure between Reaction Profiles

For defining the degree of similarity, a variety of numerical methods, such as the Pearson correlation coefficient, have been proposed. However, in this study, the Tanimoto coefficient [4),11)] was used. The Tanimoto coefficient is an index that strongly shows the relative correlation between two elements [4].

The degree of similarity $T(X, Y)$ of the reaction profile $P_x = b_{x1}b_{x2}\cdots b_{xn}$ of organism X and the reaction profile $P_y = b_{y1}b_{y2}\cdots b_{yn}$ of organism Y are defined in accordance with the Tanimoto (Jaccard) coefficient as follows.

$$T(X, Y) = \frac{N_z}{N_x + N_y - N_z} \quad (3)$$

N_x and N_y are the numbers of 1 bits in the reaction profiles P_x and P_y, respectively, and N_z is the number of common reactions in both reaction profiles P_x and P_y. By definition, $T(X, Y)$ is the number in the range 0 to 1; the closer to

1, the higher the degree of similarity between the two reaction profiles, while the closer to 0, the lower the degree of similarity between the two reaction profiles.

For example, the reaction profiles of organisms S_1, S_2, and S_3 of Fig. 1 become 1111, 1110, and 0111, respectively. Here, $T(S_1, S_2) = 3/4 = 0.75$, and $T(S_2, S_3) = 2/4 = 0.5$; thus, the similarity between the reaction profiles of S_1 and S_2 is higher than that between the reaction profiles of S_2 and S_3.

3.3 Clustering

Using the degree of similarity T as defined in Section 3.2, a dissimilarity score $D(A, B)$ was defined between reaction profiles.

$$D(A, B) = 1 - T(A, B) \quad (4)$$

Then, on the basis of the dissimilarity D, a distance matrix for all organisms was created. Clustering was performed on the dissimilarity matrix. Although there are various clustering methods, such as the group average method and the centroid method, in this study, the furthest neighbor method was used in addition to the method proposed by Hong, et al.

4. Experiments and Results

4.1 Experiments and Results

To evaluate the effectiveness of the proposed method, the metabolic networks were compared among different species.

Reaction profiles were constructed from the MetaCyc database update version 2004-09-27, 33 sequenced organisms (6 archaea, 26 bacteria, 1 eukaryotes). A list of the organisms is shown in **Table 1**. Table 1 lists organism name, abbreviation, and the number of enzyme reactions found in that organism. The number of enzyme reactions within the table is the number of enzyme reactions in the metabolic networks of each species.

The MetaCyc database was provided in flat file format, and reconstructed with MySQL.

Table 1 List of organisms used in this analysis.

	Organism	Code	Number of Reactions
Archaea			
1	*Archaeoglobus fulgidus DSM4304*	AfD	791
2	*Methanococcus jannaschii DSM2661*	MjD	693
3	*Methanobacterium thermoautotrophicum delta H*	MtD	702
4	*Pyrococcus furiosus DSM 3638*	PfD	720
5	*Thermoplasma acidophilum DSM 1728*	TaD	502
6	*Thermoplasma volcanium GSS1*	TvG	773
Bacteria			
1	*Aquifex aeolicus VF5*	AaV	687
2	*Borrelia burgdorferi B31*	BbB	473
3	*Clostridium acetobutylicum ATCC824*	CaA	896
4	*Caulobacter Crescentus*	Cc	812
5	*Campylobacter jejuni NCTC 11168*	CjN	728
6	*Campylobacter jejuni RM1221*	CjR	682
7	*Escherichia coli K-12*	EcK	1041
8	*Escherichia coli O157*	EcO	855
9	*Enterococcus faecalis V583*	EfV	817
10	*Haemophilus influenzae KW20 Rd*	HiK	836
11	*Helicobacter pylori 26695*	Hp2	542
12	*Helicobacter pylori J99*	HpJ	614
13	*Mycobacterium leprae TN*	MlT	745
14	*Neisseria meningitidis serogroup A Z2491*	NmA	790
15	*Neisseria meningitidis MC58*	NmM	800
16	*Pseudomonas aeruginosa PAO1*	PaP	1093
17	*Porphyromonas gingivalis W83*	PgW	796
18	*Streptococcus pneumoniae R6*	SpR	848
19	*Streptococcus pneumoniae TIGR4*	SpT	717
20	*Streptococcous thermophilus LMG 18311*	StL	762
21	*Streptococcus pyogenes MGAS10394*	Sy1	874
22	*Streptococcus pyogenes MGAS8232*	Sy2	768
23	*Streptococcus pyogenes SF370 serotype M1*	Sy3	868
24	*Vibrio cholerae N16961*	VcN	848
25	*Yersinia pestis CO92*	YpC	1184
26	*Yersinia pestis KIM*	YpK	946
Eukarya			
1	*Human*	Hu	1187

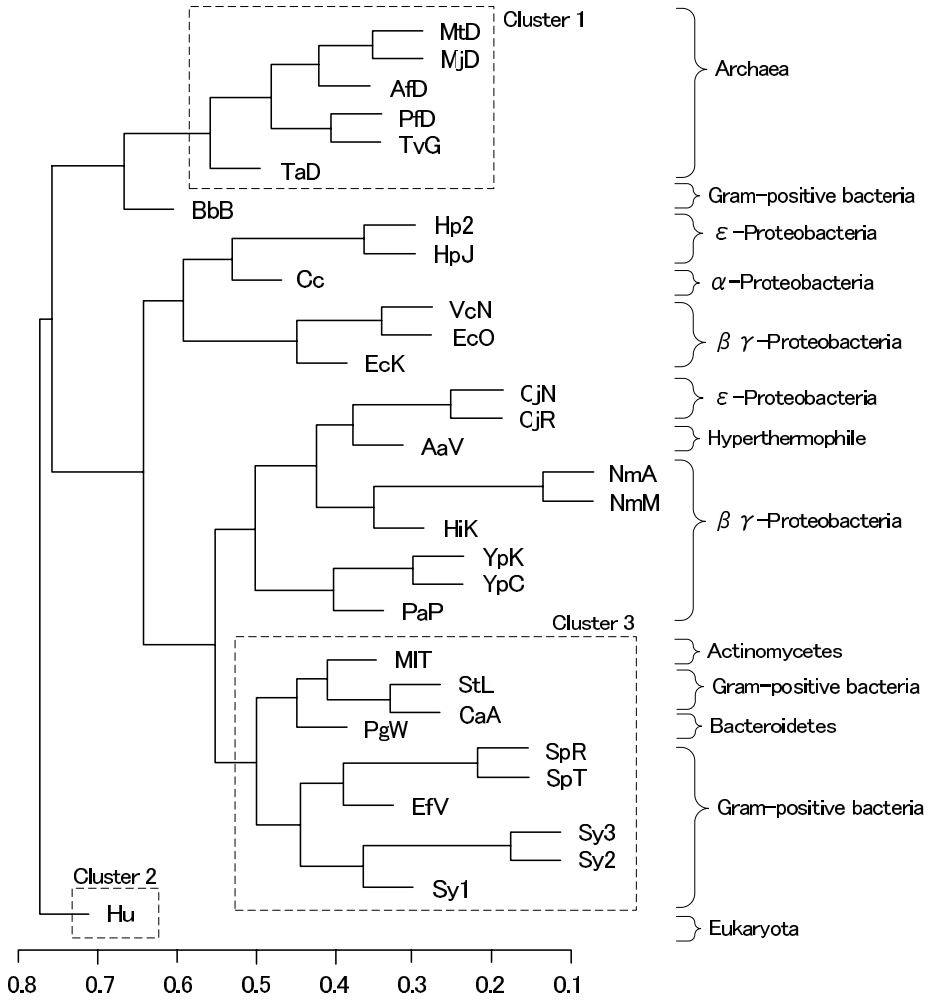

Fig. 2 Clustering result based on reaction profiles of 33 organisms.

The distance matrix is calculated using Perl. Statistical processing software R Version 2.3.0 [12] was used for clustering and for the creation of a phylogenetic tree diagram. As a result, 33 reaction profiles consisting of 3744 bits were obtained. The phylogenetic tree obtained as is shown in **Fig. 2**. The abbreviations (e.g., MtD and MjD) in Fig. 2 represent the organism names, and these correspond to their formal names shown in Table 1.

4.2 Discussion

Within the phylogenetic tree shown in Fig. 2, the six species of archaea —— *Methanobacterium thermoautotrophicum delta* (MtD), *Methanococcus jannaschii DSM2661* (MjD), *Archaeoglobus fulgidus DSM4304* (AfD), *Pyrococcus furiosus DSM 3638* (PfD), *Thermoplasma volcanium GSS1* (TvG), and *Thermoplasma acidophilum DSM 1728* (TaD) —— were classified within the same cluster (Cluster 1). Further, the sole eukaryote, *Human* (Hu), was located apart from these organisms (Cluster 2). The proposed method successfully achieved distinct separation of the archaea, the bacteria, and the eukaryote. The proposed method showed a similar tendency to the clustering results for archaea and bacteria using the method of Hong, et al. [8].

The 26 species of bacteria were widely separated; they were divided into gram-positive proteobacteria, gram-negative bacteria, and other. Nine species were classified as gram-positive bacteria, and the eight speices of gram-positive bacteria —— *Streptococcus thermophilus LMG 18311* (StL), *Clostridium acetobutylicum ATCC824* (CaA), *Streptococcus pneumoniae R6* (SpR), *Streptococcus pneumoniae TIGR4* (SpT), *Enterococ-*

cus faecalis V583 (EfV), *Streptococcus pyogenes MGAS10394* (Sy1), *Streptococcus pyogenes MGAS8232X* (Sy2), *Streptococcus pyogenes SF370 serotype M1* (Sy3) —— were classified within Cluster 3. Thus, the gram-positive bacteria, with the exception of *Borrelia burgdorferi B31* (BbB), were all found at relatively neighboring positions within Cluster 3.

BbB was clustered away from the other bacteria. The number of enzymatic reactions comprising BbB is 473, which is only abound half the number of enzymatic reactions of Sy3 (868), which has a similar number of ORFs to the other bacteria [13] listed in KEGG update version 2006-07-04 [1]. Therefore, there is a strong possibility that there are deficiencies in the metabolic reaction data for BbB, and such deficiencies are considered to have an effect on the clustering results.

In addition, I used the furthest neighbor method. However, this method is weak with regard to the outlier [14]. This weakness may have been responsible for the distant clustering of BbB. It will be necessary to evaluate which clustering technique is best for the purpose of this research by applying other clustering techniques, and comparing the results in future studies.

Moreover, it is possible that this result was due to the limitations of the method. The proposed method does not consider the situation where a similar but not identical reaction between species —— i.e., "divergent-pathways." For example, *Escherichia coli* has a transferase that synthesizes citrate and coenzyme A from oxaloacetate, acetyl-coA and H_2O (citrate synthase). However, *Chlorobium limicola* does not have the reaction. Instead of the reaction, *Chlorobium limicola* has a transferase that synthesizes citrate and coenzyme A from oxaloacetate, acetyl-coA, ADP and phosphate (ATP citrate synthase) [15].

Proteobacteria are generally classified as α-proteobacteria, ε-proteobacteria, and $\beta\gamma$-proteobacteria. With regard to the metabolic networks of these species, while organisms belonging in the same species, such as *Neisseria meningitidis serogroup A Z2491* (NmA) and *Neisseria meningitidis MC58* (NmM), were grouped closely together, there was good separation of each of the groups α, ε, and $\beta\gamma$, including *Aquifex aeolicus VF5* (Aav). These results show that, among the proteobacteria, there may exist groups with metabolic networks similar to those of gram-positive bacteria, as well as groups with metabolic networks that are not similar to those of gram-positive bacteria. This was recently reported by Zhang, et al. [16]. With regard to these organisms, further detailed investigations are required to determine where on the metabolic map their characteristics are located.

Although the proposed method closely resembles the method proposed by Yamada, et al. [4], there are several important differences, including the purposes of the respective methods and the fact that genes are used as the standard in the method of Yamada, et al.

5. Conclusions and Future Work

A method was proposed for comparing metabolic networks among different species based on their reaction profiles by considering metabolic networks as a set of enzymatic reactions. This method is a combination of the commonly used bit expression for data, the Tanimoto coefficient method, and clustering, and is relatively easy to implement. The method was applied for 33 actual species, and its validity was demonstrated.

It is anticipated that metabolic network data will become available in future for a variety of organisms, including eukaryote-specific metabolic pathways. It would be possible to obtain more knowledge considering the diversification of metabolic phenotypes using this method. Further knowledge will be acquired by comparison of phylogenetic trees between metabolic phenotypes and genomic sequences.

Nevertheless, the analysis of metabolic networks based on the method proposed here has the problem that it is markedly impacted by data deficiencies. Thus, it will be necessary to introduce (1) verification of influence that lack of data has on clustering results, (2) consideration of the divergent-pathways, (3) comparison of results according to a variety of clustering methods, and (4) statistical scale that will guarantee the reliability of results.

Acknowledgments I would like to thank Professor Matsuda and Professor Takenaka at the University of Osaka for useful suggestions. This study was, in part, supported 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, 2006.

References

1) Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M.: The KEGG resource for deciphering the genome, *Nucleic Acids Research*, 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 Research*, Vol.34, pp.D511–516 (2006).
3) Forst, C.V. and Schulten, K.: Evolution of metabolisms: a new method for the comparison of metabolic pathways using genomics information, *Journal of Computational Biology*, Vol.6, pp.343–360 (1999).
4) Yamada, T., Goto, S. and Kanehisa M.: Extraction of Phylogenetic Network Modules from Prokayrote Metabolic Pathways, *Genome Informatics*, Vol.15, No.1, pp.249–258 (2004).
5) Tohsato, Y., Matsuda, H. and Hashimoto, A.: An application of a pathways alignment method to the analysis of metabolic pathways, *Research Communications in Biochemistry, Cell and Molecular Biology*, Vol.5, pp.179–191 (2003).
6) Fitch, W.M. and Margoliash, E.: Construction of phylogenetic trees, *Science*, Vol.155, pp.279–284 (1967).
7) Feng, D.F., Cho, G. and Doolittle, R.F.: Determining divergence times with a protein clock: update and reevaluation, *Proc. Nat. Acad. Sci. USA*, Vol.94, pp.13028–13033 (1997).
8) Hong, S.H., Kim, T.Y. and Lee, S.Y.: Phylogenetic analysis based on genome-scale metabolic pathway reaction content, *Applied Microbiology and Biotechnology*, Vol.65, No.2, pp.203–210 (2004).
9) Ebenhoh, O., Handorf, T. and Heinrich, R.: A cross species comparison of metabolic network functions, *Genome Informatics*, Vol.16, No.1, pp.203–213 (2005).
10) Clemente, J.C., Satou, K. and Valiente, G.: Reconstruction of phylogenetic relationships from metabolic pathways based on the enzyme hierarchy and the gene ontology, *Genome Informatics*, Vol.16, No.2, pp.45–55 (2005).
11) Willet, P., Barnard, J.M. and Downs, G.M.: Chemical similarity searching, *Journal of Chemical Information and Computer Sciences*, Vol.38, No.6, pp.983–996 (1998).
12) http://www.rproject.org
13) http://www.genome.jp/kegg/catalog/org_list.html
14) Duda R.O., Hart P.E. and Stork D.G.: *Pattern Classification*, 2nd ed., John Wiley & Sons, Inc., New York, NY (2001).
15) Lill, U., Schreil, A. and Eggerer, H.: Isolation of enzymically active fragments formed by limited proteolysis of ATP citrate lyase, *European Journal of Biochemistry*, Vol.125, No.3, pp.645–650 (1982).
16) Zhang, Y., Li, S., Skogerbo, G., Zhang, Z., Zhu, X., Zhang, Z., Sun, S., Lu, H., Shi, B. and Chen, R.: Phylophenetic properties of metabolic pathway topologies as revealed by global analysis, *BMC Bioinformatics*, Vol.7, pp.252–264 (2006).

(Received July 10, 2006)
(Accepted September 7, 2006)

(Communicated by *Masakazu Sekijima*)

Yukako Tohsato is an Assistant Professor of 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. degree 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.