IPSJ Transactions on Bioinformatics Vol. 2 83–92 (Sep. 2009)

Original Paper

A Comparative Analysis of Metabolic Pathways Based on Metabolic Steady States

Yuta Ashida,[†1] Tomonobu Ozaki[†1]
and Takenao Ohkawa[†1]

A comparative analysis of organisms with metabolic pathways provides important information about functions within organisms. In this paper, we discuss problem of comparing organisms using partial metabolic structures that contain many biological characteristics and propose a pathway comparison method based on an elementary flux mode (EFM) — the minimal metabolic pathway that satisfies a steady state. By the extraction of the 'elementary flux mode,' we obtain biologically significant metabolic substructures. To compare metabolic pathways based on EFMs, we propose a new pseudo alignment method with a penalty based on the importance of enzymes. The distance among organisms can be calculated based on the pseudo alignment of EFMs. To confirm its effectivity, we apply the proposed method to the pathway datasets from 38 organisms. We successfully reconstructed a "three domain theory" from the aspect of the biological function. Moreover, we evaluated the results in terms of the accuracy of organism classification from the biological function and confirmed that the obtained classification was related deeply to such habitats as aerobe or anaerobe.

1. Introduction

Organisms ingest various substances such as food and other sustenance. By participating in many chemical reactions throughout the body, these substances can be transformed into the compounds and energy necessary for the continued activity of the organism. Metabolism, which implies the sum total of all such reactions, is characterized by chain reactions in which the product of one enzyme reaction becomes the substrate for other enzyme reactions. Large-scale networks that comprise these chain reactions are called metabolic pathways.

By analyzing metabolic pathways, we can obtain knowledge about the processes by which organisms have acquired various functions in the course of evolution. Furthermore, comparative analysis of metabolic pathways among different species is an important task by which significant information on the relatedness of biological functions can be understand. Conventionally, species comparisons of biological phenotypes or of genomic sequences using rRNA, etc. are employed in phylogenetic systematics. However, the former is enormously expensive to perform, and the latter is not able to reflect phenomena such as the horizontal gene transfer that occurs during evolution. Thus accurately representing the evolutionary relatedness of organisms is difficult [3]. In light of this background, it is hoped that comparative analysis of metabolic pathways will complement conventional methods of phylogenetic systematics.

In this paper, we consider the comparison of organisms using partial metabolic structures that contain many biological characteristics and propose a pathway comparison method based on an elementary flux mode (EFM) — the minimal metabolic pathway that satisfies a steady state [6),7]. Since EFMs represent metabolic substructure, they are substructures that include many biological features. Analyzing metabolic pathways with EFM enables comparative analysis based on biologically meaningful substructures.

In a previous study [1], we developed a pathway analysis method using important substructures of pathways and confirmed that important reaction structures effectively contribute to pathway comparisons. In this method, assuming that each extracted substructure, i.e. the reaction structure, has a certain size, the size of substructures is required as parameters. However, because the appropriate substructure size depends on enzymes and/or species, the size is difficult to determine in advance.

On the other hand, EFMs are uniquely extracted as biologically significant substructures. Therefore we need not prepare any size parameters in advance. Instead, to calculate similarities between EFMs of differing size, the insertion or deletion of compounds (nodes) or enzymes (edges) must be considered. The insertion or deletion of an important enzyme can be considered to represent a large difference among biological species. We therefore propose a method of pseudo alignment with a penalty based on the importance of the inserted or deleted enzyme.

†1 Kobe University

Zhu, et al. explored the structural properties of metabolic pathways based on network indices, degree distribution, and network motifs [11]. In their research, while network indices and the degree distribution were used as the overall topological features of metabolic pathway, they employed the network motifs as characteristic substructures. The employment of the network motif is deeply related to our method, since we regard EFMs as another kind of characteristic substructure in metabolic pathways. However, the measures used in their method are based on only the structural viewpoints of the metabolic pathway. The most obvious difference between their idea and ours is that we try to extract and compare substructures from biological or functional features by introducing EFMs. When the overall topology of the metabolic pathway is used for comparative analysis, both metabolic structures that include many biological features of organisms and others may be treated in the same way. In our method, however, using substructures in the steady state, and introducing pseudo alignment that reflects the importance of the enzyme, we directly extract features related to biological activity.

2. Pathway Comparison Method Based on EFM

Figure 1 shows an outline of the general flow to generate a phylogenetic tree from metabolic pathways. First EFMs are extracted from the given metabolic pathways, and then the extracted EFMs are aligned using pseudo alignment with penalties. The distance between the given species is calculated based on the alignment results. Finally, a phylogenetic tree is generated using a distance matrix.

2.1 Elementary Flux Mode

The metabolic flow rate dynamically changes in organisms. However, a long-term view of metabolic properties can be obtained by extracting pathways that constitute steady states [6]. EFM is intuitively defined as a minimal pathway that satisfies steady states and that becomes non-steady when divided further. When a metabolic pathway satisfies $\mathbf{0} = \mathbf{Nr}$, that pathway is a steady state. Here, $\mathbf{N}$ represents an $m \times q$ matrix, where m is the number of compounds and q is the number of enzyme reactions. In addition, $\mathbf{r}$ is a $q \times 1$ vector and $\mathbf{0}$ is an $m \times 1$ null vector.

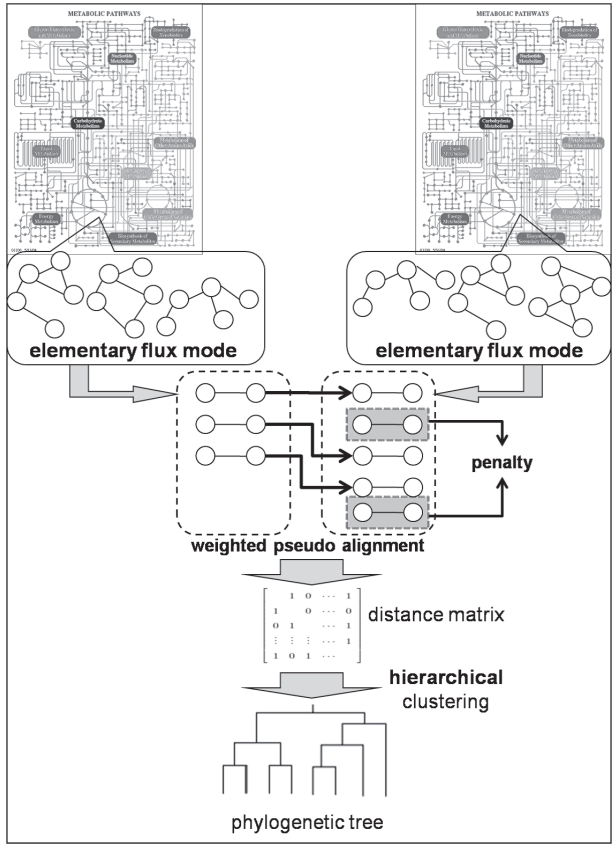

Fig. 1 Phylogenetic tree generation by pathway comparison.

Examples of EFM from *E.coli* purine metabolism are shown in **Table 1**. EFMs of different sizes were extracted.

2.2 Penalty Based on Enzyme Importance

Rahman, et al. defined load points (LP) and choke points (CP) as two types of important enzymes or compounds that exist in a metabolic pathway [5]. An LP is an enzyme through which many of the shortest paths of a metabolic pathway traverse. In other words, after obtaining the shortest paths for all combinations

Table 1 Example of Elementary Flux Mode from *E.coli* purine metabolism. Strings that begin with 'R' such as 'R00259' are *R number* which represents reaction, and numeric values represent reaction rate. 'External' means reaction to external pathways.

mode 1	1 R01229 -1 R01131 1 R01863 -1 R02147 -1 R01132 1 R01228
mode 2	1 R01229 -1 R01131 1 R01863 -1 R02147 -1 R02142 1 R01228 -1 R01130 1 R01768
mode 3	-1 external 1 R01858 -1 R00190 1 R02090 1 R01131 -1 R00332 -1 R01863 2 R00439 -1 R01969 1 R02557 -2 R00437 1 R02147 1 R00127 1 R00376 1 R00124 -1 R00430 2 R01857 1 R02088 1 R01132 -1 R01137 -1 R01968 -1 R01547 -1 R01228 -1 R00375
mode 4	1 external -1 R01229 2 R00441 -2 R00435 2 R01858 1 R00190 -1 R02090 1 R00332 -1 R01138 1 R01969 -1 R02557 -1 R00127 -1 R00376 -1 R00124 -1 R00430 1 R01857 -1 R02088 1 R01968 1 R01547 1 R00375
mode 5	-1 external 1 R01858 1 R00190 -1 R02090 -1 R01131 1 R00332 -1 R01138 1 R01863 -1 R00439 1 R01969 -1 R02557 1 R00437 -1 R02147 -1 R00127 -1 R00376 -1 R02088 -1 R01132 1 R01968 1 R01547 1 R01228 1 R00375

of enzymes, enzymes with high load values (LV) are determined by the number of shortest paths traversing through each enzyme and the number of enzymes to which each enzyme is adjacent. Enzymes with high LVs are considered LPs.

In contrast, a CP is an enzyme that, when lost, significantly negatively impacts the maintenance of life processes. If there are many LPs on the shortest path of a metabolic pathway, that pathway is considered to play a major role in metabolism. An enzyme through which many such shortest paths traverse is particularly important for an organism, and its loss adversely affects the maintenance of life processes. More formally, the choke value (CV) of a particular enzyme represents the sum of the LVs for all enzymes located on a shortest path traversing through that enzyme, and enzymes with high CVs become CPs.

In this paper, we used this CV to determine the degree of importance of an enzyme (penalty) at the alignment process. The CVs for all enzymes on a metabolic pathway are calculated and sorted in descending order, and then the importance of an enzyme is evaluated using the following expression (1):

$$penalty(e) = \frac{rank(e)}{|\mathbf{E}|}. \tag{1}$$

Here, $\mathbf{E}$ is the set of enzymes contained in an organism's metabolic pathway, and $rank(e)$ is the CV-based rank order of enzyme e, which is an element of $\mathbf{E}$. CVs are obtained for each organism, however, depending on the organism in which the enzyme is included; values may differ greatly. For example, Halobacterium sp., from the archaea species, has a maximum value of 9,433, while in Homo sapiens, a eukaryote species, the maximum value is 59,238. This is because the CV value in an organism with many enzymes becomes larger on average based on the nature of the CV calculation method. Therefore, rather than using the CV itself, we use its rank order among the enzymes within the same organism to calculate the penalty.

2.3 Calculation of Distance from a Pseudo Alignment with Penalty

We give the pathway alignment algorithm that can calculate the distance between two organisms. The algorithm is described in a similar manner as the method proposed by Clemente, et al.[2].

An organism is represented as a set of EFMs. The similarity of organisms X and Y is defined as

$$sim(X,Y) = \frac{1}{|\mathbf{P} \cup \mathbf{Q}|}\left(|\mathbf{P} \cap \mathbf{Q}| + \sum_{P \in \mathbf{P} \setminus \mathbf{Q}} \max_{Q \in \mathbf{Q}} sim(P,Q) + \sum_{Q \in \mathbf{Q} \setminus \mathbf{P}} \max_{P \in \mathbf{P}} sim(P,Q)\right), \tag{2}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are sets of EFMs extracted from the pathways of X and Y, respectively.

To calculate the similarity between EFMs, we propose a pseudo alignment with penalty in which the importance of an enzyme is considered a penalty. Similarity $sim(P,Q)$ between EFMs P and Q is defined as

$$sim(P,Q) = \frac{\prod_{S \in \mathbf{S} \setminus \mathbf{R}} penalty(S)}{|\mathbf{R}|} \left(|\mathbf{R} \cap \mathbf{S}| + \sum_{R \in \mathbf{R} \setminus \mathbf{S}} \max_{S \in \mathbf{S}} sim(R,S) \right), \quad \left(|R| \leq |S| \right), \tag{3}$$

where $\mathbf{R}$ and $\mathbf{S}$ represent the sets of enzyme reactions that constitute EFMs P and Q, respectively, and *penalty* represents the penalty value ($0 < penalty \leq 1$) based on the importance of the enzyme reaction. For each element of set $\mathbf{R}$ of enzyme reactions, the most similar reaction is selected from $\mathbf{S}$, and the similarities are added. Subsequently, the similarity value is reduced by multiplying the penalty value of enzyme reactions among $\mathbf{S}$ that do not correspond to any of the reactions of $\mathbf{R}$. This is based on the idea that in the course of evolution, the loss of the enzyme itself represents a serious evolutionary change.

In addition, the similarity of enzyme reactions R and S is defined as

$$sim(R,S) = \frac{1-\alpha}{|\mathbf{C} \cup \mathbf{D}|} |\mathbf{C} \cap \mathbf{D}| + \frac{\alpha}{|\mathbf{E} \cup \mathbf{F}|} \left(|\mathbf{E} \cap \mathbf{F}| + \sum_{E \in \mathbf{E} \setminus \mathbf{F}} \max_{F \in \mathbf{F}} sim(E,F) + \sum_{F \in \mathbf{F} \setminus \mathbf{E}} \max_{E \in \mathbf{E}} sim(E,F) \right), \tag{4}$$

where $\mathbf{C}$ and $\mathbf{D}$ represent sets of compounds and $\mathbf{E}$ and $\mathbf{F}$ represent sets of enzymes in enzyme reactions R and S, respectively. $\alpha (0 \leq \alpha \leq 1)$ represents the weight of the similarity of the compounds against the similarity of the enzymes. Here, we set $\alpha = 0.5$ because we believe that both compounds and enzymes represent the features of biological reactions. If α becomes higher, the modification process of compounds in the metabolic steady state will be extracted. In contrast, if α becomes lower, the function of enzymes that consist of the metabolic steady state are reflected. The similarity of compounds is set to either 1 or 0. Hierarchical similarity [9] is used for similarity between enzymes.

The distance between biological species X and Y can be calculated using the similarity defined in expression (2) as

$$1 - sim(X,Y). \tag{5}$$

3. Experiment

3.1 Outline

To confirm the effectiveness of our proposed method, we conducted experiments on the comparative analysis of metabolic pathways. Taking pairs that could be formed from all combinations of the target biological species, distances were calculated using the proposed method and phylogenetic trees were created by clustering.

All of the metabolic pathways used in the experiments were obtained from KEGG. A CellNetAnalyzer tool converted the pathway data in KEGG to chemical reaction matrices and also extracted the EFMs using the chemical reaction matrices. R ver. 2.4.1 was used for clustering.

The 38 biological species analyzed in this study are shown in **Table 2**. The species name is abbreviated in three letters. In addition, 'Domain' represents the three domain types of archaea, eubacteria, and eukaryote [10]. Although many kinds of organisms were analyzed and the results stored in databases such as KEGG, we believe that the 38 organisms are sufficient for the evaluation of domain classification because they were chosen from each domain.

In general, the constructed phylogenetic tree is evaluated through a comparison with a molecular gene-based phylogenetic tree. However, since the genes represent the blueprints for those parts that together comprise the biological functions and do not represent the biological functions themselves, the molecular phylogenetic trees constructed by gene-comparisons do not necessarily constitute classifications of the organisms based on biological function. Erroneous classifications based on horizontal gene transfer can also occur. Therefore, from the perspective of the phylogenetic comparisons of biological functions, it is not always appropriate to consider gene-based molecular phylogenetic trees as accurate.

We therefore assessed the effectiveness of the proposed method not only using comparisons with molecular phylogenetic trees but also by other databases related to biological function to determine the quality of functional classifications. We utilized Genome Seek, a genome database from Genamics ⋆1, as reference data for

⋆1 http://genamics.com/genomes/index.htm

Table 2 Thirty eight organisms included in phylogenetic analysis.

Abbr.	Organism	Abbr.	Organism
Domain: Archaea		eco	Escherichia coli K-12 MG1655
hal	Halobacterium sp.	fnu	Fusobacterium nucleatum
mac	Methanosarcina acetivorans	hin	Haemophilus influenzae
mja	Methanococcus jannaschii	hpj	Helicobacter pylori J99
mma	Methanosarcina mazei	lla	Lactococcus lactis
mth	Methanobacterium thermoautotrophicum	mlo	Mesorhizobium loti
pab	Pyrococcus abyssi	nma	Neisseria meningitidis serogroup A
pai	Pyrobaculum aerophilum	oih	Oceanobacillus iheyensis
pfu	Pyrococcus furiosus	pmu	Pasteurella multocida
pho	Pyrococcus horikoshii	sco	Streptomyces coelicolor
sso	Sulfolobus solfataricus	sme	Sinorhizobium meliloti
sto	Sulfolobus tokodaii	spy	Streptococcus pyogenes
tac	Thermoplasma acidophilum	stm	Salmonella typhimurium
tvo	Thermoplasma volcanium	sty	Salmonella typhi
Domain: Eubacteria		tte	Thermoanaerobacter tengcongensis
atc	Agrobacterium tumefaciens C58 Cereon	Domain: Eukaryote	
bme	Brucella melitensis	cel	Caenorhabditis elegans
cac	Clostridium acetobutylicum	dme	Drosophila melanogaster
cpe	Clostridium perfringens	hsa	Homo sapiens
cte	Chlorobium tepidum	sce	Saccharomyces cerevisiae
dra	Deinococcus radiodurans		

the classification of biological functions and evaluated the classification results by our method based on habitat. That is because the habitat of organisms is one of the most effective factors for biological functions.

Genamics provides classification of biological functions as a gateway to genome data. We therefore added interpretations to the constructed phylogenetic trees based on the functions provided by Genamics and then performed evaluations based on these interpretations. For the eukaryotes and the 'mja' and 'mth' archaea, we did not provide an interpretation since Genamics did not list the biological functions.

3.2 Results

The phylogenetic tree constructed by the proposed method with the domain labels is shown in **Fig. 2**. In taxonomy, organisms are first divided into broad groupings of archaea, eubacteria, and eukaryotes known as domains, based on the three domain hypothesis proposed by Woese, et al.[10]. Each of the domains has biological characteristics that differ greatly from each other; such a biological organism domain classification system is well established. As shown in Fig. 2, most of the organisms in our phylogenetic tree are divided into three domains: archaea, eubacteria, and eukaryotes. This result indicates that a functional perspective of organisms also supports the three domain hypothesis.

While there are numerous hypotheses in the field of biology concerning evolutionary processes in archaea, eubacteria, and eukaryotes, the most prevalent hypothesis states that archaea and eubacteria were the first to differentiate, after which a species of archaea evolved into eukaryotes[3]. In Fig. 2, however, the eukaryote cluster is seen adjacent to the eubacteria cluster, which is contrary to the above hypothesis. Many species of archaea live in peculiar environments and have relatively simple functionalities. Eubacteria, however, lives in a variety of environments and therefore possesses complex metabolic functions. Eukaryotes possess the most complex metabolic functions due to the advanced structural organization of their cells into organisms. Therefore, from the perspective of metabolic functions, eubacteria are indeed properly positioned near eukaryotes, indicating that a correct understanding of the metabolic functions of organisms was achieved.

In addition, to compare the obtained phylogenetic tree with that derived from gene sequences, we measured its second cousin similarity[8] with the phylogenetic

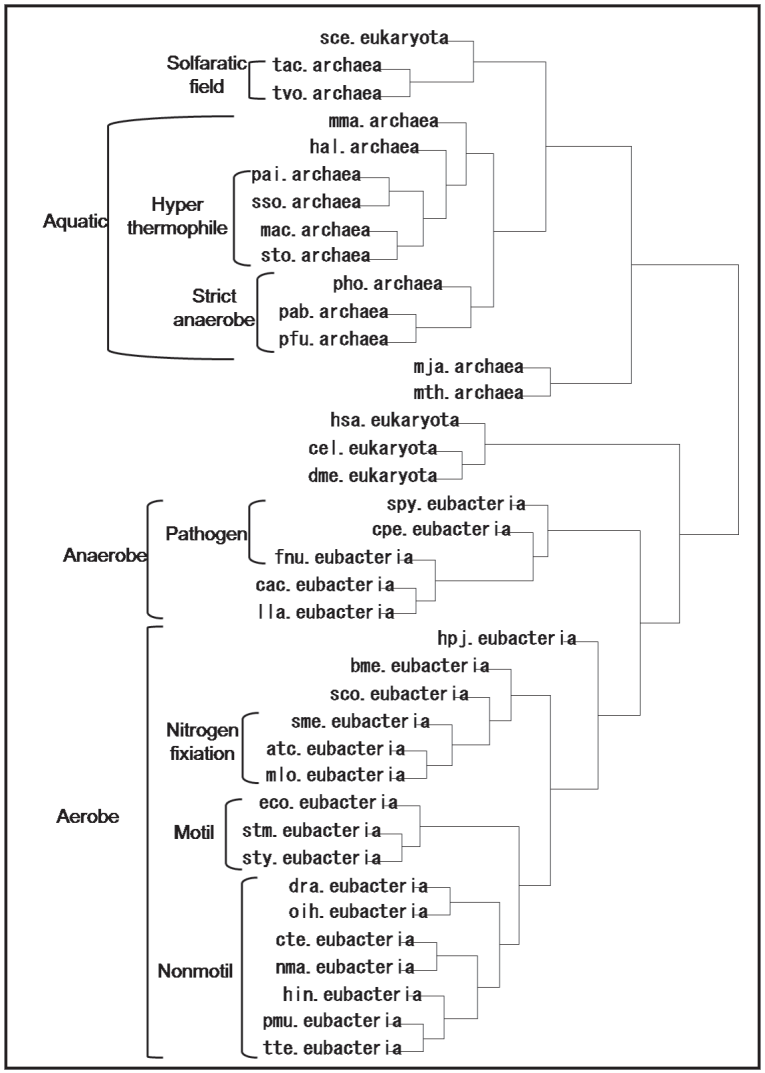

Fig. 2 Phylognetic tree by proposed method.

Table 3 Classification accuracy.

(a) proposed method

	precision	recall	F-value
anaerobic	1.00	0.71	0.83
aerobic	0.83	1.00	0.93

(b) without penalty

	precision	recall	F-value
anaerobic	0.54	0.86	0.67
aerobic	0.90	0.54	0.75

tree obtained from NCBI. Second cousin similarity, which is an evaluation measure generally used for evaluating the similarities between phylogenetic trees, is obtained by comparing partial trees within six degrees of separation and then repeating the process. However, this measure typically evaluates differentiation conditions near branch ends rather than at early stages, and similarity tends to weaken as the size of the phylogenetic tree increases. The similarity with NCBI based on second cousin similarity was 0.19. This result is relatively high based on metabolic pathway comparisons across the three domains. However, since agreement with the molecular phylogenetic tree is not the goal of the proposed method, it is not necessarily superior to all previous studies on this point.

Furthermore, the constructed phylogenetic trees are examined from the perspective of biological function. They can be classified almost completely based on their habitat; organisms with such similar habitat environment as oceanic, aerobic, anaerobic etc. were classified into the same clusters. Within these clusters, they were further divided by such biological functions as pathogenicity, dynamicity, immobility, etc.

To quantitatively evaluate the classification accuracy by biological function, we conducted an experiment using two types of organisms: aerobic and anaerobic. In this experiment, the classifications were evaluated for 7 anaerobic and 14 aerobic species from eubacteria. The results are summarized in **Table 3** (a). For the anaerobic class, the precision was 1.00, the recall was 0.71, and the F-value was 0.83. For the aerobic class, the precision was 0.88, the recall was 1.00, and the F-value was 0.93; classification by biological function is quite accurate.

3.3 Comparative Experiment

To confirm the effectiveness of CP-based pseudo alignment with penalty, a phylogenetic tree was constructed from the same biological species using an alignment method proposed by Clemente, et al.[2]. Note that this alignment method does not consider penalties. The results are shown in **Fig. 3**.

A number of biological species are classified into different domains. Taking the 'mja' and 'mth' archaea as examples, only these organisms are classified into a completely different cluster. This is likely due to the fact that the pathway data for 'mja' and 'mth' are still incomplete, and thus sufficient EFMs could not be obtained. In fact, for archaea it was possible to extract 300 EFMs on average, but the EFMs obtained from these two archaea were less than 100. In contrast, these organisms were correctly classified into the archaea cluster in our method.

The above suggests that by penalizing changes in important enzymes, the characteristics of an organism's basic metabolic functions are strengthened.

We further evaluated the classifications by biological function. In Fig. 3, organisms with hyperthermophilic properties are divided into two clusters. Moreover, within the eubacteria cluster, the aerobic organism cluster is further divided into two more clusters and the anaerobic organism cluster is classified as occurring in between these clusters. This shows that classification failed in relation to the fundamental function of respiration, which calculates the energy necessary for survival.

In evaluating the quantitative precision of classification, as shown in Table 3 (b), in the anaerobic class the precision was 0.54, the recall was 0.86, and the F-value was 0.67. In the aerobic class, the precision was 0.90, the recall was 0.64, and the F-value was 0.75. High precision was obtained even with comparisons by EFM only, but a low recall value was obtained. Using the proposed method and focusing on the important enzymes resulted in a higher level of precision, and the functional relatedness of organisms could be reproduced more accurately.

3.4 Comparison with Related Work

We compare the proposed method with a scheme using graph kernel [4] (kernel-based method), which is also based on structural information. The phylogenetic tree generated by the kernel-based method is shown in **Fig. 4**, which is depicted based on the original figure in Ref. 4). The kernel-based method completely

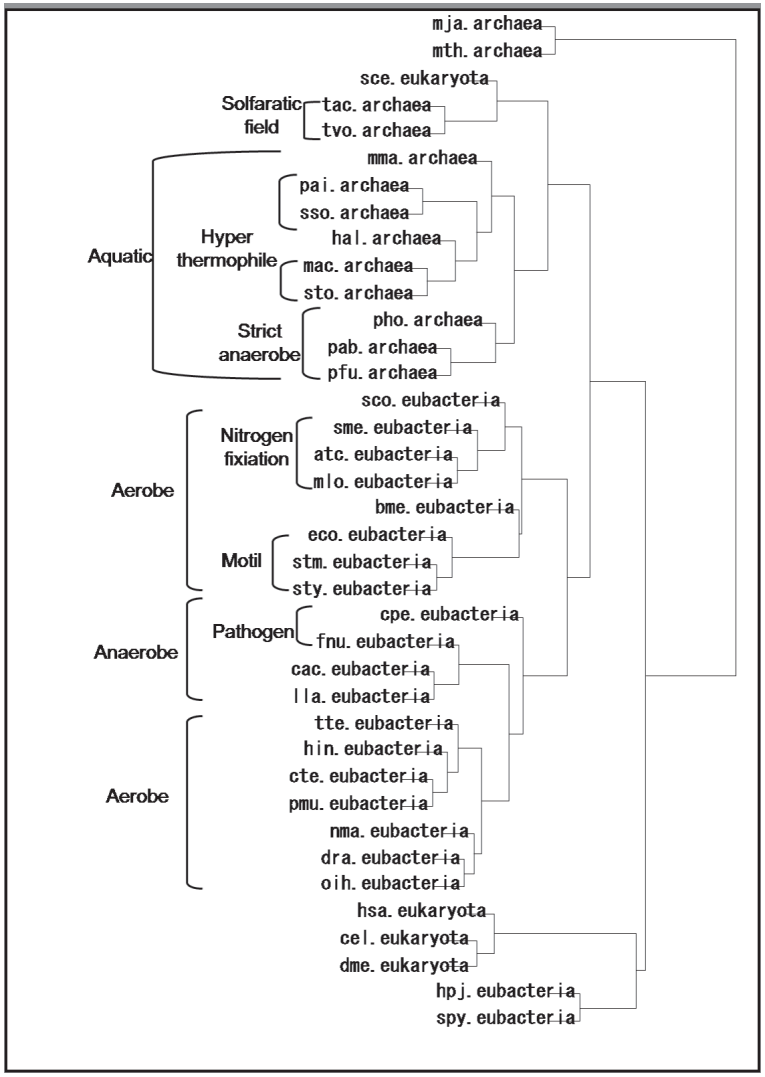

Fig. 3 Phylogenetic tree with elementary flux mode and biological function.

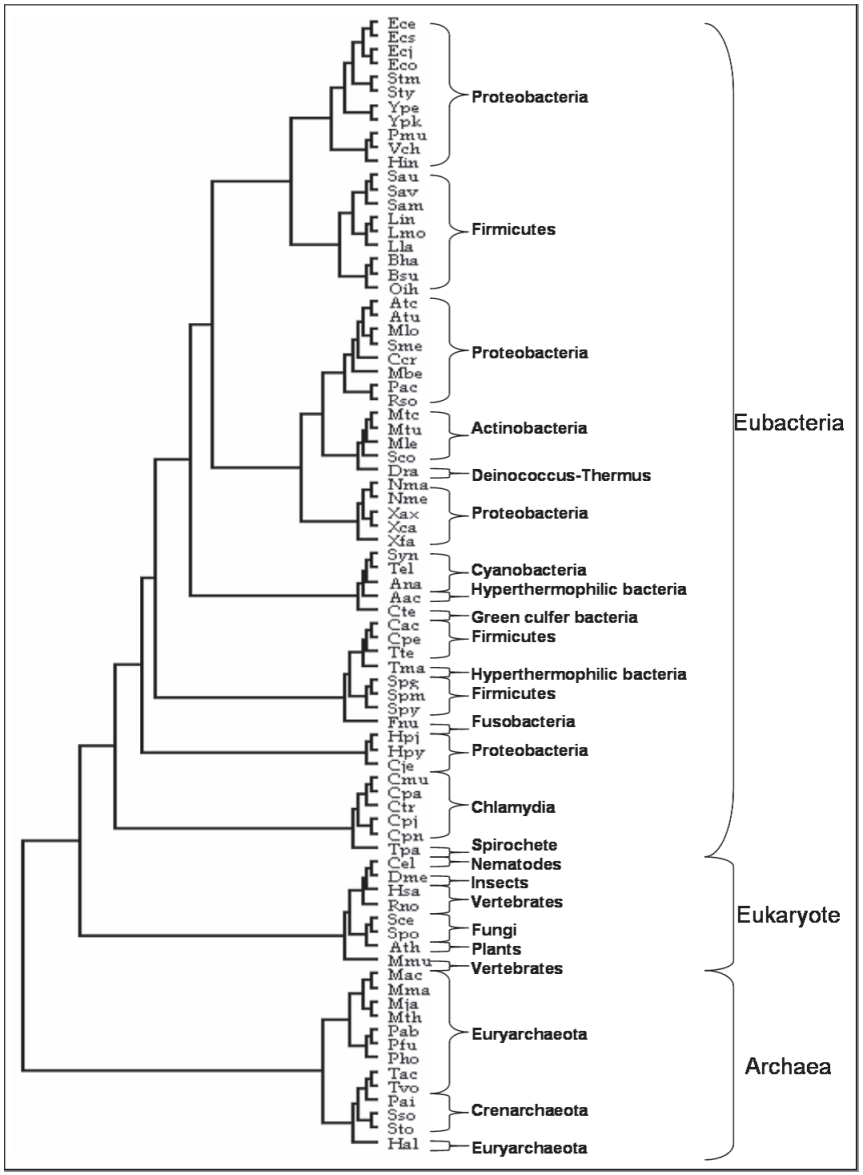

Fig. 4 Phylogenetic tree by kernel-based method (Modified from Ref. 4)).

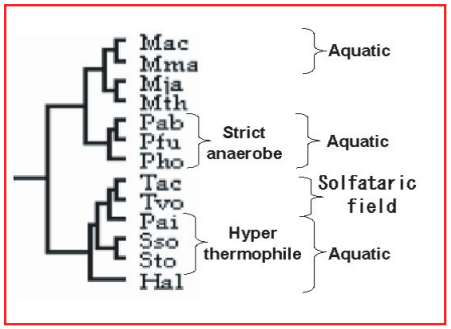

Fig. 5 Phylogenetic tree in archaea by kernel-based method (Modified from Ref. 4)).

classifies three domains in more organisms in comparison with our experiments. Moreover, the method reproduces a three domain theory from the viewpoint of function as well as the proposed method does.

The proposed method is slightly inferior to the method in the domain classification in all domains. However, the second cousin similarity to the NCBI taxonomy of the phylogenetic tree based on the kernel-based method is 0.17. Thus, the proposed method resembles a conventional phylogeny on classification in the domain and outperforms the kernel-based method because the proposed technique supplements conventional phylogeny.

Next, we compare classification results with the kernel-based method from the viewpoint of biological function. Eukaryote and eubacteria, part of the target organisms used in the experiment reported by Oh, et al., are quite different from our experiment. On the other hand, archaea are almost the same. Therefore, we only evaluate the classification for the archaea domain. **Figure 5** shows part of the classification of archaea by the kernel-based method and interpretations based on function from Genamics. In Fig. 5, organisms that have the same functional properties such as 'strict anaerobe' and 'hyper thermophile' are classified into the same cluster. However, the kernel-based method fails to classify such ecotypes as 'aquatic' and 'solfataric field.'

The proposed method classifies organisms based on ecotypes first, and then by functional property. In this point, the proposed method realized classification reflecting biological function.

3.5 Discussion

From the above experiments, we reproduced a three domain hypothesis and classified organisms by biological function by comparing EFMs based on a pseudo alignment method that penalizes enzymes based on their importance in metabolic pathways. The three domains represent the most fundamental groupings of organisms, and a classification system that deviates from this biological premise is considered inferior to one that does not. The phylogenetic tree constructed by the proposed method largely conforms to this biological premise and can accurately represent the functional relatedness of organisms.

There are exceptions, however, where certain portions of the classification system deviate substantially from the fundamental biological premise. In the phylogenetic tree depicted in Fig. 2, the eukaryote 'sce' is contained in the archaea cluster, which contradicts the fundamental premise; so sufficient functional characteristics were not obtained. However, in all comparative experiments based on EFM, including pilot studies, 'sce' was positioned near the archaea. 'Sce', which is a species of budding yeast, is a single-celled organism. In addition, it shares properties such as the ability to ferment alcohol through anaerobic metabolism with the archaea group, which inhabits peculiar environments. Since this raises the possibility that there are portions functionally similar to the archaea, a more detailed investigation is necessary.

4. Conclusion

In this paper, using EFM — the minimal metabolic pathway in the steady state — we proposed a pseudo alignment method with penalty based on enzyme importance levels. The characteristics of the proposed method are summarized as follows:

- By focusing on EFMs, we can compare metabolic pathways based on the metabolic structure in the most stable state of the organism.
- A pseudo alignment with penalty is introduced to generate a phylogenetic tree based on substructure comparisons reflecting enzyme importance.

A comparative analysis of metabolic pathways was performed in 38 biological species. Using the proposed method, the three domain hypothesis was reproduced from a biological function perspective. In addition, a biological function-based evaluation revealed that classification deeply related to such habitat environments as aerobic or anaerobic could be reproduced using the proposed method.

In future studies, we must apply the proposed method to many more biological species and conduct species comparison experiments. While the number of biological species used in this paper is sufficient to classify organisms across the three domains, it remains impossible to conduct detailed comparisons because the biological species used differed greatly from those used in related studies. Therefore we must clearly demonstrate the effectiveness of this method through experimentation using many more biological species. In addition, we hope to examine the influence of the biological features of organisms by analyzing the contribution of EFM based on the proposed method.

References

1) Ashida, Y., Ozaki, T. and Ohkawa, T.: Reaction Structure Profile: A comparative analysis of metabolic pathways based on important substructures, *IPSJ Transaction on Bioinformatics*, Vol.49, No.5, pp.36–45 (2008).
2) Clemente, J.C., Satou, K. and Valiente, G.: Phylogenetic reconstruction from non-genomic data, *Bioinformatics*, Vol.23, pp.e110–e115 (2006).
3) Feng, D.F., Cho, G. and Doolittle, R.F.: Determining divergence times with a protein clock: update and reevalution, *Proc. National Academy of Science of USA*, Vol.94, pp.13028–13033 (1997).
4) Oh, S.J., Joung, J.G., Chang, J.H. and Zhang, B.T.: Construction of phylogenetic trees by kernel-based comparative analysis of metabolic networks, *BMC Bioinformatics*, Vol.7, No.284, pp.1471–2105 (2006).
5) Rahman, S.A. and Schomburg, D.: Observing local and global properties of metabolic pathways: 'load points' and 'choke points' in the metabolic networks, *Bioinformatics*, Vol.22, No.14, pp.1767–1774 (2006).
6) Schuster, S., Dandekar, T. and Fell, D.A.: Detection of elementary flux modes in biochemical networks: A promising tool for pathway analysis and metabolic engineering, *Trends Biotechnology*, Vol.17, pp.53–60 (1999).
7) Schuster, S. and Hilgetag, C.: On elementary flux modes in biochemical reaction systems st steady state, *Journal of Biological Systems*, Vol.2, No.2, pp.165–182 (1994).
8) Shasha, D., Wang, J.T.L. and Zhang, S.: Unordered tree mining with applications to phylogeny, *Proc. 20th International Conference on Data Engineering*, pp.708–719 (2004).
9) Tohsato, Y., Matsuda, H. and Hashimoto, A.: A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy, *Proc. 8th International*

Conference on Intelligent Systems for Molecular Biology, pp.376–383 (2000).
10) Woose, C.R., Kandler, O. and Wheelis, M.L.: Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya, *Proc. National Academy of Sciences*, Vol.87, pp.4576–4579 (1990).
11) Zhu, D. and Qin, Z.S.: Structural comparison of metabolic networks in selected single cell organisms, *BMC Bioinformatics*, Vol.6, No.8, pp.1471–2105 (2005).

(Received March 27, 2009)
(Accepted June 15, 2009)
(Released September 24, 2009)

(Communicated by *Yukako Tohsato*)

Yuta Ashida received his B.E. and M.E. degrees from Kobe University in 2007 and 2009 respectively. His major research interests include bioinformatics and graph mining.

Tomonobu Ozaki received his Ph.D. in Media and Governance from Keio University in 2002. Now he is an assistant professor of Organization of Advance Science and Technology, Kobe University. He is a member of JSAI.

Takenao Ohkawa received his B.E., M.E., and Ph.D. degrees from Osaka University in 1986, 1988, and 1992, respectively. He is currently a professor at the Department of Computer Science and Systems Engineering, Graduate School of Engineering, Kobe University. His research interests include intelligent software and bioinformatics. He is a member of the IEEE, the Information Processing Society in Japan, the Institute of Electronics, Information, and Communication Engineers, the Institute of Electrical Engineers in Japan, and the Japanese Society for Artificial Intelligence.