TopEVM: Using Co-occurrence and Topology Patterns of Enzymes in Metabolic Networks to Construct Phylogenetic Trees

Tingting Zhou[1,2], Keith C.C. Chan[2], and Zhenghua Wang[1]

[1] National Laboratory for Paralleling and Distributed Processing,
National University of Defense Technology, Changsha, Hunan, 410073, P.R. China
Tingting.Zhou@live.com, zhhwang188@sina.com
[2] Department of computing, The Hong Kong Polytechnic University, Hong Kong, China
cskcchan@inet.polyu.edu.hk

Abstract. Network-based phylogenetic analysis typically involves representing metabolic networks as graphs and analyzing the characteristics of vertex sets using set theoretic measures. Such approaches, however, fail to take into account the structural characteristics of graphs. In this paper we propose a new pattern recognition technique, *TopEVM*, to help representing metabolic networks as weighted vectors. We assign weights according to co-occurrence patterns and topology patterns of enzymes, where the former are determined in a manner similar to the *Tf-Idf* approach used in document clustering, and the latter are determined using the degree centrality of enzymes. By comparing the weighted vectors of organisms, we determine the evolutionary distances and construct the phylogenetic trees. The resulting *TopEVM* trees are compared to the previous *NCE* trees with the NCBI Taxonomy trees as reference. It shows that *TopEVM* can construct trees much closer to the NCBI Taxonomy trees than the previous *NCE* methods.

Keywords: *TopEVM*, phylogenetic analysis, metabolic network, co-occurrence pattern, document clustering, topology pattern, degree centrality, evolutionary distance.

1 Introduction

The objective of phylogenetic analysis is to reconstruct the evolutionary relationship among different species and to display them in a tree-structured model called a *phylogenetic tree* [1]. Applications include the design of new drugs and the reconstruction of the history of infectious diseases [2]. Most previous research [3] in this area has been based on sequence alignment but these sequence-based approaches are easily influenced by horizontal gene transfer (HGT) [4, 5]. An alternative to this is network-based phylogenetics analysis, which compares the homogeneous biological networks of organisms. They often make use of metabolic networks and take the quantified difference between these networks as the evolutionary distance.

M. Chetty, A. Ngom, and S. Ahmad (Eds.): PRIB 2008, LNBI 5265, pp. 225–236, 2008.
© Springer-Verlag Berlin Heidelberg 2008

A metabolic network is a hierarchical, graph-represented abstract of an actual metabolism. Composed of thousands of metabolites, enzymes, reactions and the relationships among them, global metabolic networks are too large and complicated to be compared element by element. So, for comparison purposes, the vertex sets of graphs, rather than the entire graph, is common to use. In such cases, the evolutionary distances are determined by applying set theoretic measures [6-10]. For example, Aguilar et al [11] treat the organisms as enzyme sets from the view of metabolism, building a binary vector for each organism according to the presence or absence of the enzymatic functions. Using the *NCE* (Number of Common Enzymes) method and a normalized *Hamming distance*, they construct phylogenetic trees by creating a distance matrix for each metabolic class. Forst et al [8] construct 'clean' metabolite-reaction bipartite graphs to represent metabolic networks. Using the *Jaccard distance* as the evolutionary distance measure, they construct the distance matrix by taking organisms as reaction sets. Tohsato [7] consider metabolic networks as enzymatic reaction sets. Also using the *Jaccard distance*, she determines the evolutionary distance matrix and constructed phylogenetic trees. One drawback of such set-theoretic methods is that they do not usually take into account the edge information, and therefore they do not have enough topological characteristics for the network comparison, especially the topological importance of vertices [9, 12, 13].

In this paper we propose a new pattern recognition technique, *TopEVM*, for use in phylogenetic analysis. *TopEVM* avoids a common drawback of set-theoretic methods in that it takes account of the structural characteristics of graphs by representing metabolic networks as weighted vectors. We assign the weights based on the co-occurrence and topology patterns of enzymes in organisms, where the co-occurrence patterns are determined using a method similar to the *Tf-Idf* approach in the document clustering and the enzyme topology patterns are determined according to the degree of centrality of enzymes. By comparing the weighted vectors of organisms, we determine the evolutionary distance matrices for the construction of phylogenetic trees. Comparing to the previous set-theoretic methods, *TopEVM* can produce phylogenetic trees closer to the taxonomy trees of NCBI.

The remainder of this paper is organized as follows. Section 2 elaborates the *TopEVM* approach. Section 3 describes our experiments and results. Section 4 provides conclusion and outlines directions for the future work.

2 *TopEVM*: Constructing Phylogenetic Trees by Using Enzyme Co-occurrence and Topology Patterns

In this section we describe the operation of the *TopEVM* approach, which proposes the use of a frequency weighting scheme and a topological vector. This approach proceeds from the observation that it is possible to regard the construction of species trees in phylogeny as similar to the process of distance-based clustering of organisms which may in turn be seen as analogous to document clustering, with an organism as a document and an enzyme as a term. This allows us to apply feature extraction approaches and the hierarchical clustering methods to the construction of phylogenetic trees.

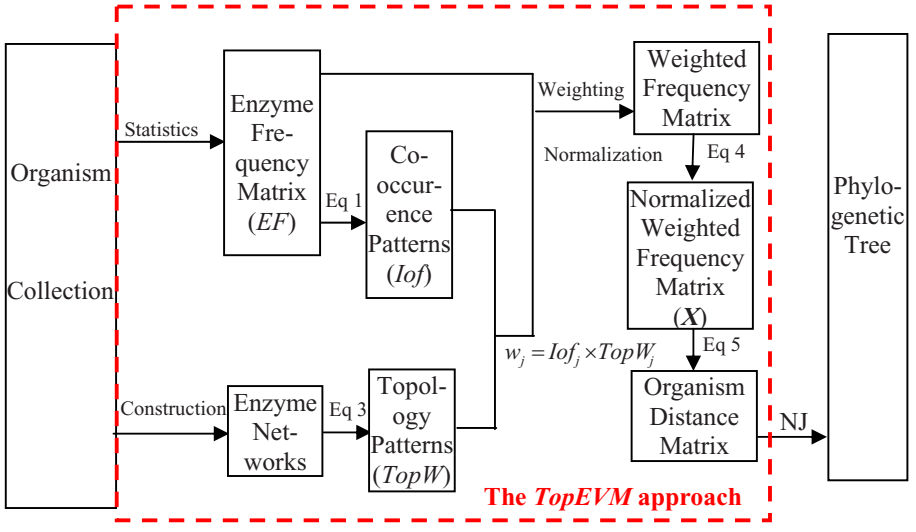

Fig. 1. The *TopEVM* approach

The first step in the *TopEVM* approach is to set up a matrix to record how frequently an enzyme occurs in a given collection of organisms, and extract enzyme co-occurrence patterns according to this matrix. By analogy with the *Tf-Idf* weighting scheme used in document clustering, we define *Inverse organism frequency* (*Iof*), a weight vector, to extract enzyme co-occurrence patterns. In the second step *TopEVM* uses a topological weight vector, *TopW*, to extract topologically important enzymes, the enzyme topology patterns, for use as features. This is done by representing metabolic networks as enzyme graphs and then counting the degree centrality, one measure of topological importance, of enzymes. The next step is to normalize the enzyme co-occurrence and topology weighting schemes. These are then used to convert the original frequency matrix into a new matrix in which rows denote the final Topology-weighted Enzyme Vector Model (*TopEVM*) of organisms. Finally a distance matrix is established by comparing the *TopEVM* of organisms with *Soergel Distance* as the distance measure. The distance matrix is used to construct the phylogenetic trees by use of some distance-based clustering approach, e.g., Neighbor Joining (NJ) method. Figure 1 shows the flow of the entire procedure.

2.1 Inverse Organism Frequency: Extracting Enzyme Co-occurrence Patterns

The first step to extract the co-occurrence patterns of enzymes is to set up a matrix to record how frequently an enzyme occurs in a given collection of organisms. For this purpose, we define *Enzyme Vector Space* to denote the organism-enzyme frequency matrix and *Enzyme Vector Model* to denote the organisms as enzyme frequency vectors. It should be noted that we make two assumptions in the definitions. First, we assume that

enzymes are arranged in the ascending order of their EC numbers[1] and the order is kept constant. Second, we assume that organisms in the collection are arranged in an arbitrary but constant order.

Enzyme Vector Space. Let O be a collection of m organisms o_i, $i = 1..m$. Let E be all the n enzymes e_j, $j = 1..n$, of at least one organism in O. The *Enzyme Vector Space* O^{mn} is defined through the organism-enzyme frequency matrix, EF, where ef_{ij} is the frequency of the j^{th} enzyme, $j = 1..n$, in the i^{th} organism.

Enzyme Vector Model. Given an enzyme vector space O^{mn} and its organism-enzyme frequency matrix, EF, the enzyme frequency vector for the i^{th} organism o_i is defined as the i^{th} row of EF. We call this representation of organisms as *Enzyme Vector Model*.

Document clustering generally assumes that the total term frequency is not always indicative of a term's information content. To account for this disparity, the *Inverse document frequency* (*Idf*) weighting scheme is often applied [14]. We find a similar situation when we compare the enzyme frequency vectors of organisms. That is to say, the frequency of an enzyme appeared in all the organisms cannot be assumed to indicate its information content. To deal with this, we apply a weighting scheme in this study, which is similar to the *Idf* weighting scheme. We call this the *Inverse organism frequency* (*Iof*) weighting scheme and define it as follows.

Organism Frequency. Given an enzyme vector space O^{mn} and its organism-enzyme frequency matrix EF, the *Organism Frequency* (of_j) of a given enzyme e_j, $j = 1..n$, is defined as the number of organisms that contain the enzyme e_j.

Inverse Organism Frequency. Given a enzyme vector space O^{mn} and its organism-enzyme frequency matrix EF, the *Inverse organism frequency* (Iof_j) of a given enzyme e_j, $j = 1..n$, is defined as the logarithm of the quotient of dividing the total organism number by its organism frequency (of_j). That is,

$$Iof_j = \log\left(m \Big/ \sum_i I_{(ef_{ij}>0)} \right) \tag{1}$$

where $I_{(\cdot)}$ denotes the indicator function, $I_{(cond)} = \begin{cases} 1 & if.cond.is.fulfilled \\ 0 & otherwise \end{cases}$, and $\sum_i I_{(ef_{ij}>0)}$ is the organism frequency of e_j, namely of_j.

Once Iof_j has been assigned to enzyme e_j, the original frequency of enzyme e_j in organism o_i, namely ef_{ij}, can be transformed into a new weighted frequency ef_{ij}',

$$ef_{ij}' = Iof_j \cdot ef_{ij} \tag{2}$$

Since the *Iof* weighting scheme gives lower weights to the enzymes found in a large number of organisms and higher weights to those found in fewer organisms, the *Iof* weights emphasize organism-specific enzymes.

[1] The EC (Enzyme Commission) number is a numerical classification scheme for enzymes.

2.2 Topology Weight: Extracting Enzyme Topology Patterns

Some enzymes in a metabolic network will have a higher average connectivity than others [15]. On the assumption that this higher average connectivity represents topological important information, we define *Topology Vector Space* to denote the organism-enzyme topology matrix and *Topology weight* (*TopW*) weighting scheme to select more strongly connected enzymes.

Topology Vector Space. Let $\boldsymbol{R}$ be a collection of m metabolic networks. r_i is constructed for the i^{th} organism o_i, $i = 1..m$. Let $\boldsymbol{E}$ be the collection of all the n enzymes e_j, $j = 1..n$, which are contained by at least one metabolic network in $\boldsymbol{R}$. The Topology Vector Space $\boldsymbol{R}^{mn}$ is defined through the organism-enzyme topology matrix, $\boldsymbol{T}$, where t_{ij} is the topological importance of the j^{th} enzyme in the metabolic network of the i^{th} organism, $j = 1..n$.

In this study, the degree centrality, the number of direct neighbors of a node [16], is regarded as the measure of the node's topological importance. In order to distinguish the absent enzymes from the present enzymes with degree as '0', we assign the degree centrality of the absent enzymes as '-1' in the topology matrix $\boldsymbol{T}$.

Topology Weight. Given a topology vector space R^{mn} and its organism-enzyme topology matrix T, the *Topology weight* ($TopW_j$) of the given enzyme e_j, $j = 1..n$, is defined as:

$$TopW_j = \sum_i \left(t_{ij} \cdot I_{(t_{ij} \geq 0)}\right) \Big/ \sum_i I_{(t_{ij} \geq 0)} \tag{2}$$

where $I_{(\cdot)}$ is an indicator function defined as in Eq 1, and $\sum_i I_{(t_{ij}>0)}$ is the total number of metabolic networks in $\boldsymbol{R}$ containing enzyme e_j.

The *TopW* weighting scheme gives higher weights to the enzymes with higher average degree, which strengthens the importance of more highly-connected enzymes.

2.3 Normalization: Eliminating the Influence of Vector Length on Distance

The difference of the vector length can influence the calculation of the distance between vectors. *Iof* and *TopW* weighting schemes help select the 'important' enzymes as the features of organisms, but result in the organism vectors with different lengths. Therefore, it is necessary to normalize the weighted organism vectors before calculating the distance between them.

Let $\boldsymbol{X}$ denotes the weighted and normalized organism-enzyme frequency matrix, where the element x_{ij}, $i = 1..m$, $j = 1..n$, is given by

$$x_{ij} = \frac{w_j \times f_{ij}}{\sqrt{\sum_k \left(w_k \times f_{ik}\right)^2}}, \quad w_j = Iof_j \times TopW_j \tag{3}$$

The rows in $\boldsymbol{X}$ are the final representative of organism vectors.

2.4 Soergel Distance: Calculating the Evolutionary Distance

Soergel distance is one of the distance measures which are commonly used to calculate the evolutionary distance, a crucial measure of the similarity of organism vectors. It has the advantages that its range is limited to 0~1 and it obeys the triangular inequality [17].

Suppose $\boldsymbol{X_A}$ and $\boldsymbol{X_B}$ are two vectors of equal length n, the Soergel Distance between them is defined as:

$$D_{A,B} = \frac{\sum_{j=1}^{n} \left| x_{jA} - x_{jB} \right|}{\sum_{j=1}^{n} \max\left(x_{jA}, x_{jB} \right)} \tag{4}$$

The distance matrix can be established by calculating the Soergel distance between organism vectors pair-wisely. It can be used to construct the phylogenetic trees using a suitable distance-based clustering algorithm, e.g. Neighbor-joining (*NJ*) method.

3 Experiments and Results

In this study, enzyme and reaction data are obtained from the database created by Ma and Zeng [18]. The Ma and Zeng database consists of five tables: *reaction*, *enzyme*, *react*, *connect* and *organism*, and contains 3663 enzymes and 107 organisms (8 Eukaryotes, 83 Bacteria and 16 Archaea) in total. We acquire enzyme frequency information from *enzyme*, and construct the enzyme graphs for each organism from *enzyme* and *reaction*.

Although the *TopEVM* approach is capable of dealing with large collections of organisms, for the sake of concision, in this explanation we select only eight organisms: *rno*, *mmu*, *afu*, *mja*, *nme*, *hin*, *lin* and *bsu*.

Table 1 lists the details of these 8 organisms: their ID in KEGG database (*KEGG ID*), their full name(*Organism*), the Kingdom they belong to (*Kingdom*)□their ID in NCBI Taxonomy [19] (*NCBI Tax Id*), and the number of enzymes they contain (N_E).

Table 1. The details of the 8 organisms

KEGG ID	Organism	Kingdom	NCBI Tax ID	N_E
rno	Rattus norvegicus	Eukaryota	487	416
mmu	Mus musculus	Eukaryota	727	470
afu	Archaeoglobus fulgidus	Archaea	1423	277
mja	Methanococcus jannaschii	Archaea	1642	244
nme	Neisseria meningitides	ProteoBacteria	2190	369
hin	Haemophilus influenzae	ProteoBacteria	2234	386
lin	Listeria innocua	Bacteria Firmicute	10090	388
bsu	Bacillus subtilis	Bacteria Firmicute	10116	504

3.1 Distribution of the *Iof* Weights

According to the definitions in Section 2.1, omitting the absent enzymes in all of the 8 organisms, we calculate the *Iof* weight vector of length 1063. The *Iof* weights values distribute over 8 points. Table 2 displays the value of N_E along the *Iof* Weights. Nearly 30% enzymes have the highest *Iof* weights. 2.1% enzymes will be neglected for their *Iof* importance are 0. Moreover, there are around 60% enzymes with the *Iof* value over 1, and 40% between 0 and 1. Since the *Iof* weighting scheme gives lower weights to the enzymes occurring in a large number of organisms, it comes that the lower the *Iof* value is, the more organisms the enzyme spreads in. This observation also confirms the conclusion of Liu et al [13]. That is, most of the enzymes occur in several organisms they prefer, while only few enzymes occur in most of the studied organisms.

Table 2. The statistics of the number of enzymes along the *Iof* weights

***Iof* Weight**	**2.08**	**1.39**	**0.98**	**0.69**	**0.47**	**0.29**	**0.13**	**0**
Num Enzymes	318	310	102	118	71	85	36	23
Percentage(%)	30	29.1	9.6	11.1	6.7	8.0	3.4	2.1

3.2 Distribution of the *TopW* Weights

In order to calculate the *TopW* weights, we represent the enzyme networks upon the following principles: vertices denote individual enzymes and arcs denote the relationships between them; if one enzyme's product is the substrate of another enzyme, then there's an arc directed from the former enzyme to the latter. The bidirectional arc is replaced by two individual arcs with opposite direction.

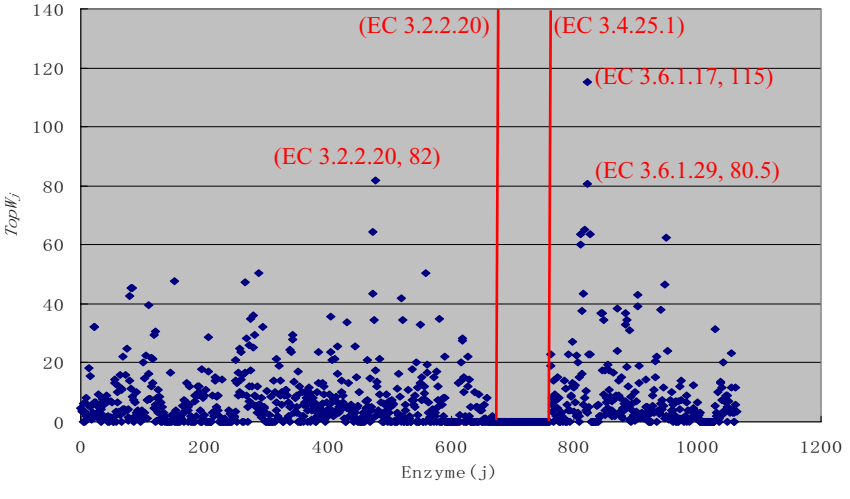

Fig. 2. Distribution of the *TopW* weights. The x-coordinates denote the ordered enzymes. The y-coordinates denote the corresponding *TopW* weight value of the enzyme. The coordinates of the top 3 enzymes are marked on the figure. The continuous range between two vertical dash lines denotes the range of enzymes whose *TopW* weight is 0.

Table 3. The enzymes with the top 8 *TopW* value

Order	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Enzyme	3.6.1.17	2.7.4.10	3.6.1.29	3.6.1.15	2.7.4.6	3.6.1.3	3.6.3.1	4.6.1.1
TopW	115	82	80.5	65	64.25	63.5	63.5	62.3

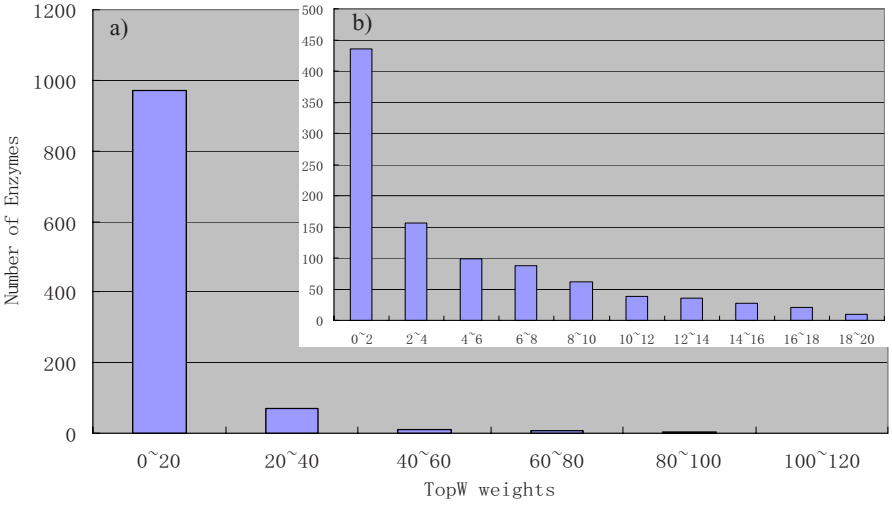

Fig. 3. Distribution of the number of enzymes along the *TopW* weights It contains two diagrams. Fig 3a) shows the distribution of enzymes all over the range of the *TopW* weights. Fig 3b) expands the *TopW* weight range of 0~20.

We construct the *TopW* weight vector for the ordered enzyme array on the basis of the definitions in Section 2.2. Fig.2 shows the distribution of the *TopW* weight values. Most of the *TopW* weights are small but several are very big. For example, *dinucleoside tetraphosphatase* (EC 3.6.1.17)'s *TopW* weight is 115, *AMP phosphotransferase* (EC 2.7.4.10)'s is 82, and *bis(5'-adenosyl)-triphosphatase* (EC 3.6.1.29)'s is 80.5. This observation indicates that as a topology pattern, the *TopW* weights of few enzymes are high, while that of most enzymes are low. We also notice that from *EC 3.2.2.20* to *EC 3.4.25.1*, there is a gap in which 110 enzymes have *TopW* values as 0. They are part of glycosyl hydrolases (*EC 3.2.*-.-), and all of the hydrolases acting on ether bonds (*EC 3.3.*-.-) as well as peptide bonds (*EC 3.4.*-.-). It is mostly due to either the large absence of the enzymes or their possible isolation.

Table 3 displays the enzymes with the top 8 *TopW* weights. It shows that the *hydrolases acting on acid anhydrides* (*EC 3.6.*-.-) have more connection, which means *hydrolases* may be more topologically important than the enzymes with other function.

Fig 3 shows the distribution of the number of enzymes along the *TopW* range. It can be seen in Fig 3a) that more than 90% of enzymes are found within the *TopW* range of

0~20 (the first tallest bar). Fig 3b) expands the first bar of Fig 3a), and displays its details of the distribution. As is shown in Fig 3b), there are 436 enzymes in the *TopW* range of 0~2. The number is nearly 41% of the total number of the enzymes. It indicates that most of nodes have very low connectivity but a handful of nodes (hubs) have much higher connectivity in the constructed enzyme networks. This result is in accordance with the scale free property of metabolic networks, and shows that enzymes become topologically different during evolution [15].

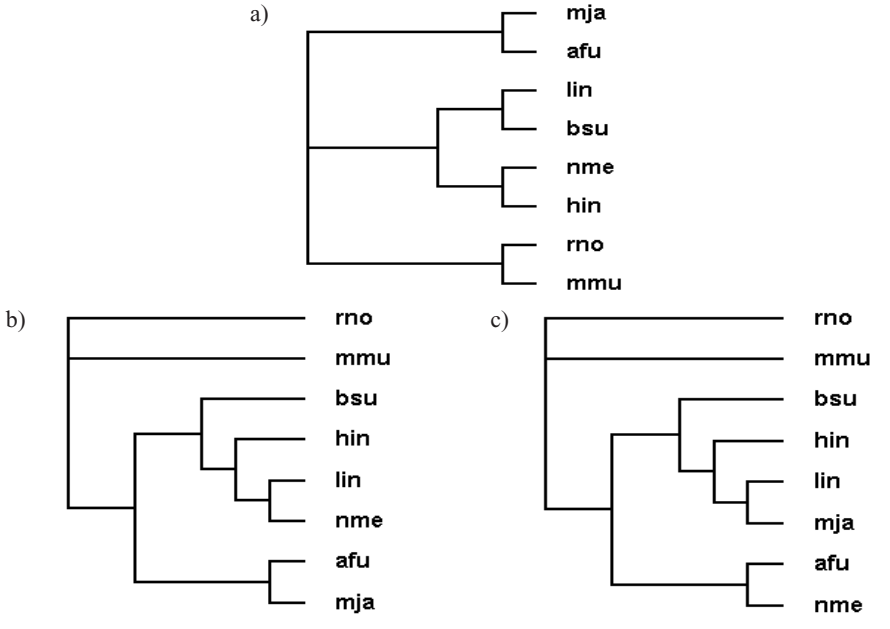

Fig. 4. The comparison of a) the NCBI tree, b) the *TopEVM* tree, and c) the *NCE* tree

3.3 Construction and Evaluation of *TopEVM* Phylogenetic Tree

We calculate the distance for each organism pair by use of Soergel distance, and obtain the distance matrix (Table 4) for constructing the phylogenetic tree. With the help of the Phylip [20] package, we use *NJ* (Neighbor Joining) method to do the construction. The resulting tree is rootless, which is displayed as Fig 5b) by use of TreeView [21]. We also obtain the phylogenetic tree from NCBI Taxonomy (Fig 5a) as the reference, and construct trees using the *NCE* method (Fig 5c) for evaluation.

As is shown in the *TopEVM* tree, the two Archaea *afu* and *mja* are grouped together undoubtedly, which is in line with the taxonomy from NCBI, and so do the two Eukaryotes *rno* and *mmu*. In the *NCE* tree, although *rno* and *mmu* are grouped together, the 4 Bacteria and 2 Archaea are mixed up.

We use *TOPD/FMTS* [22] to evaluate the similarities of trees. This software is complemented with a randomization analysis to test the null hypothesis that the similarity

Table 4. The resulting distance matrix of the 8 organisms upon *TopEVM*

	mmu	*rno*	*afu*	*nme*	*mja*	*hin*	*lin*	*bsu*
mmu	0.0000	0.1079	0.7201	0.7224	0.6754	0.6905	0.7468	0.6790
rno	0.1709	0.0000	0.7613	0.7793	0.7520	0.7613	0.8020	0.7073
afu	0.7201	0.7613	0.0000	0.2737	0.4409	0.4911	0.6667	0.5287
nme	0.7224	0.7793	0.2737	0.0000	0.4471	0.3951	0.6266	0.5617
mja	0.6754	0.7520	0.4409	0.4471	0.0000	0.2395	0.5382	0.3533
hin	0.6905	0.7613	0.4911	0.3951	0.2395	0.0000	0.4438	0.3129
lin	0.7468	0.8020	0.6667	0.6266	0.5382	0.4438	0.0000	0.2486
bsu	0.6790	0.7073	0.5287	0.5617	0.3533	0.3129	0.2486	0.0000

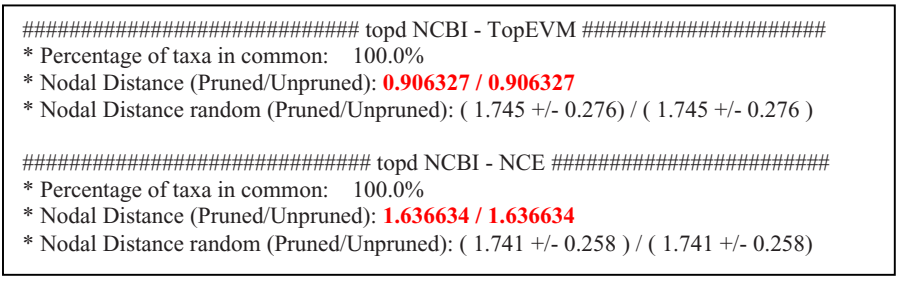

Fig. 5. The comparison result of the *TopEVM* tree and the *NCE* tree with the NCBI tree as reference

between two trees is not better than chance. With the NCBI tree as the reference, the comparison result of the *TopEVM* tree and the *NCE* tree is showed in Fig 6, which shows the *TopEVM* tree is closer to the NCBI tree with a less Nodal Distance [23] as 0.9.

4 Conclusion and Future Work

This paper proposes a new pattern recognition technique, *TopEVM*, which represents the metabolic networks as weighted vectors. By calculating the distances among these weighted vectors, evolutionary distance matrices are determined for the construction of phylogenetic trees. Comparing to the previous set-theoretic methods, our *TopEVM* method results a phylogenetic tree closer to the taxonomy tree of NCBI, which shows *TopEVM* can be a very useful approach for the network-based phylogenetic analysis.

Nevertheless, our experiments so far have considered only the *Tf-Idf* weighting scheme to integrate enzyme's frequency content. It is hard to say that there is no other weighting scheme which is more suitable. Besides, among the extensive topological indices, we only considered the degree centrality; we would also like to consider more topological information for improving our model further.

Acknowledgments. We thank Dr. Hong-Wu Ma and Prof. Dr. An-Ping Zeng for sharing their revised metabolic database. We also thank Dr. Puigbò for the direction on the use of TOPD/FMTS. This work is supported by the National Natural Science Foundation of China (6077 3021).

References

1. Husmeier, D.: Introduction to Statistical Phylogenetics. In: Husmeier, D., Richard Dybowski, R., Roberts, S. (eds.) Probabilistic Modeling in Bioinformatics and Medical Informatics. Springer, Heidelberg (2006)
2. Holmes, E.C., et al.: Using Phylogenetic Trees to Reconstruct the History of Infectious Disease Epidemics. In: Harvey, P. (ed.) New Uses for New Phylogenies. Oxford University Press, Oxford (1996)
3. Ludwig, W., Schleifer, K.: Bacterial Phylogeny Based on 16S and 23S rRNA Sequence Analysis. FEMS Microbiol Rev. 15(2-3), 155–173 (1994)
4. Wolf, Y.I., et al.: Genome Trees and the Tree of Life. Trends in Genetics 18(9), 472–479 (2002)
5. Pal, C., Papp, B., Lercher, M.J.: Adaptive Evolution of Bacterial Metabolic Networks by Horizontal Gene Transfer. Nature Genetics 37(12), 1372–1375 (2005)
6. Ebenhöh, O., Handorf, T., Heinrich, R.: A Cross Species Comparison of Metabolic Network Functions. Genome Informatics 16(1), 203–213 (2005)
7. Tohsato, Y.: A Method for Species Comparison of Metabolic Networks Using Reaction Profile. IPSJ Digital Courier 2(0), 685–690 (2006)
8. Forst, C.V., Flamm, C., Hofacker, I.L., Stadler, P.F.: Algebraic Comparison of Metabolic Networks, Phylogenetic Inference, and Metabolic Innovation. BMC Bioinformatics 7(1), 67–78 (2006)
9. Zhou, T., Chan, C., Pan, Y., Wang, Z.: An Approach for Determining Evolutionary Distance in Network-Based Phylogenetic Analysis. In: Măndoiu, I., Sunderraman, R., Zelikovsky, A. (eds.) ISBRA 2008. LNCS (LNBI), vol. 4983. Springer, Heidelberg (2008)
10. Ma, H.W., Zeng, A.P.: Phylogenetic Comparison of Metabolic Capacities of Organisms at Genome Level. Molecular Phylogenetics and Evolution 31(1), 204–213 (2004)
11. Aguilar, D., et al.: Analysis of Phenetic Trees Based on Metabolic Capabilities Across the Three Domains of Life. Journal of Molecular Biology 340(3), 491–512 (2004)
12. Zhu, D., Qin, Z.S.: Structural Comparison of Metabolic Networks in Selected Single Cell Organisms. BMC Bioinformatics 6(8) (2005)
13. Liu, W., Lin, W., Davis, A., Jordan, F., Yang, H., Hwang, M.: A Network Perspective on the Topological Importance of Enzymes and Their Phylogenetic Conservation. BMC Bioinformatics 8(121) (2007)
14. Aizawa, A.: An Information-theoretic Perspective of Tf-idf Measures. Information Processing and Management 39(1), 45–65 (2003)
15. Light, S., Kraulis, P., Elofsson, A.: Preferential Attachment in the Evolution of Metabolic networks. BMC Genomics 6(1), 159 (2005)
16. Aittokallio, T., Schwikowski, B.: Graph-based Methods for Analyzing Networks in Cell Biology. Briefings in Bioinformatics 7(3), 243 (2006)
17. Willett, P., Barnard, J.M., Downs, G.M.: Chemical Similarity Searching. J. Chem. Inf. Comput. Sci. 38(6), 983–996 (1998)

18. Ma, H.W., Zeng, A.P.: Reconstruction of Metabolic Networks from Genome Data and Analysis of Their Global Structure for Various Organisms. Bioinformatics 19(2), 270–277 (2003)
19. NCBI Taxonomy, http://www.ncbi.nlm.nih.gov/Taxonomy/
20. Phylip, http://evolution.genetics.washington.edu/phylip.html
21. TreeView, http://taxonomy.zoology.gla.ac.uk/rod/treeview.html
22. Puigbo, P., Garcia-Vallve, S., McInerney, J.O.: TOPD/FMTS: A New Software to Compare Phylogenetic Trees. Bioinformatics 23(12), 1556 (2007)
23. Bluis, J., Shin, D.G.: Nodal Distance Algorithm: Calculating a Phylogenetic Tree Comparison Metric. In: Proceedings of third IEEE Symposium on Bioinformatics and Bioengineering (2003)