A New Distance-based Approach for Phylogenetic Analysis of Protein Sequences

Shengli Zhang, Tianming Wang

***Abstract*—With the availability of ever-increasing gene and protein sequence data across a large number of species, reconstruction of phylogenetic trees to reveal the evolutionary relationship among those species becomes more and more important. In this paper, we take the physicochemical properties of amino acids into account and introduce the protein feature sequences into phylogenetic analysis by using the Bhattacharyya distance. The phylogenetic trees on the two data sets have illustrated that the proposed approach performs equally well as the other methods do and is more efficient than some of the methods. So our method may be used to complement phylogenetic analysis.**

***Index Terms*—Bioinformatics, Protein, Feature sequences, Characteristic vector, Bhattacharyya distance, Phylogenetic tree**

I. Introduction

Development of the nucleotide and protein sequencing technology have resulted in an explosive growth in the number of known DNA and protein sequences, it has raised many fundamental and challenging questions to modern biology. The elucidation of the evolutionary history of different species is a major concern to biological science. Early approaches to deal with it were mainly based on the alignment of a gene or protein sequence, but traditional alignment methods are computationally intensive and meaningless to whole genome comparison because each genome has its own genes and gene order. Accordingly, there is an urgent need to develop new phylogenetic methods utilizing the ever-increasing genome data.

Some researchers explored many methods for phylogenetic analysis, for instance, distance methods, maximal parsimony methods, maximum likelihood methods and Bayesian methods[1–8], each of which has its own range of applicability. Biologists and researchers are always trying to develop efficient methods for complex phylogenetic analysis[9–23]. Zhang et al. proposed to use gene content to measure the distance, which did not perform efficiently when the gene content of the organisms under study are very similar[24]. Yu et al. used the multiplicative model to analyze character string frequencies and derive phylogenies, where each protein was represented by a composition vector[25]. This method operates only on protein primary structures and can be applied to all genome sequences that are accompanied by nearly complete sets of predicted coding regions. Information theory is also used for phylogenetic analysis[26]. For biological sequences, the physicochemical properties of nucleic acids or amino acids are crucial factors that affect their structures or functions. The mutation of nucleic acids or amino acids is not disorderly and unsystematic. As is well known, purine is prone to be substituted by purine and pyrimidine is prone to be substituted by pyrimidine in the evolutionary process of DNA sequences. And the functions and structures of proteins are highly conserved in the evolutionary process. Liu et al. have proposed that the hydropathy profile can detect more distantly evolutionary relationships[27]. Motivated by their work, in this paper, we propose to take the protein feature sequences into account for phylogenetic analysis for distantly related proteins.

Traditional alignment method is much empirical to select or create a sequence alignment score matrix, the difference of which may affect alignment results tremendously. To overcome the problem, during the last twenty years, several alignment-free techniques for phylogenetic analysis have been developed. The Bhattacharyya distance is a theoretical distance measure between two probability distributions[28, 29]. It also has the desirable properties of being computationally simple. In this paper, we study using the classification-based Bhattacharyya distance measure to analyze the phylogeny of proteins.

II. Materials and methods

A. Protein feature sequences

Protein primary structures are linear amino acids sequences. They play an important role in determining the $3D$ structures and functions of proteins because of the physicochemical properties of amino acids. Twenty different kinds of amino acids can be divided into four classes: non-polar, negative polar, uncharged polar and positive polar in the detailed HP model[30]. The eight residues designating the non-polar class are: ALA, ILE, LEU, MET, PHE, PRO, TRP, VAL; the two residues designating the negative polar class are: ASP, GLU; the seven residues designating the uncharged polar class are: ASN, CYS, GLN, GLY, SER, THR, TYR; and the remaining three residues: ARG, HIS, LYS designate positive polar class.

Accordingly, protein primary structures can be transformed into their corresponding feature sequences. For better display, we define feature sequences for protein primary structures according to the following rule:

$$R(S(i)) = \begin{cases} 0 & S(i) = A, I, L, M, F, P, W, V \\ 1 & S(i) = D, E \\ 2 & S(i) = N, C, Q, G, S, T, Y \\ 3 & S(i) = R, H, K. \end{cases}$$

where $S(i)$ represents the ith letter in protein primary structure S and $R(S(i))$ represents the substitution for $S(i)$. From the above transformation we can see that protein feature sequence is defined in the finite set{0,1,2,3}, this four digits represent the two-double tendency of the corresponding amino acids, so protein feature sequence is the protein letter description based on two-double tendency.

For example, for the protein primary structure *S=VFFPDETGTGSYHMRWGSTQQCQVFEGLDEQQ*, its feature sequence is *R(S)=00001122222230302222222001201122*.

Since the protein feature sequence can detect more distantly evolutionary relationships, so we will, in the following section, make use of protein feature sequence to help analyze the phylogeny of distantly related proteins. We will see how much the protein feature sequences can tell us about phylogeny.

B. Characteristic vectors of protein feature sequences

Given a protein feature sequence of length L, let $N(a_1a_2\ldots a_k)$ be the occurrences of k-word $a_1a_2\ldots a_k$ observed in sequence, where a_i is one of the four digits $0, 1, 2$ or 3 and k is the word length($1 \le k \le L$). The frequency of $a_1a_2\ldots a_k$ is defined by

$$f(a_1a_2\ldots a_k) = N(a_1a_2\ldots a_k)/(L-k+1) \quad \text{(II.1)}$$

Mutations happen in a more or less random manner at the molecular level, while selections shape the direction of evolution. From the perspective of molecular evolution, k-word frequency may reflect both the results of random mutation and selective evolution. One should subtract the random background from the simple counting result in order to highlight the contribution of selective evolution[31–33]. Here, we estimate the probability of random background by using the zero-order Markov model:

$$f^0(a_1a_2\ldots a_k) = f(a_1)f(a_2)\ldots f(a_k) \quad \text{(II.2)}$$

where k ranges from 2 to L.

In this work, we collect

$$\alpha(a_1a_2\ldots a_k) = \begin{cases} \frac{f(a_1a_2\ldots a_k) - f^0(a_1a_2\ldots a_k)}{f^0(a_1a_2\ldots a_k)}, \\ \qquad f^0(a_1a_2\ldots a_k) \neq 0; \\ 0, \\ \qquad f^0(a_1a_2\ldots a_k) = 0. \end{cases} \quad \text{(II.3)}$$

for all possible words $a_1a_2\ldots a_k$ as components to constitute the characteristic vectors of protein feature sequence, which can discriminate between sequences from different species.

For a fixed k, there are total 4^k distinct k-words to be considered. Putting these k-words in a fixed order, we can get a 4^k-dimension vector denoted by $(\alpha_1, \alpha_2, ..., \alpha_{4^k})$, where α_i means the characteristic of the ith k-word. We can construct a k-word characteristic vector $A_k = (\alpha_1^A, \alpha_2^A, ..., \alpha_{4^k}^A)$ for sequence A and likewise $B_k = (\alpha_1^B, \alpha_2^B, ..., \alpha_{4^k}^B)$ for sequence B. The selection of word length k is very important to capture rich evolutionary information of protein sequence. From the view of information theory, word length reflects the balance between noise and information–some information may be lost and noise will dominate if overshort words or relatively long words are considered. We will find the balance point of noise and information in phylogenetic analysis of protein sequences.

C. Bhattacharyya distance

The Bhattacharyya distance is covered in many texts on statistical pattern recognition. In statistics, the Bhattacharyya distance measures the similarity of two discrete probability distributions. It is normally used to measure the separability of classes in classification.

The Bhattacharyya distance is a measure of divergence. It can be defined formally as follows. Let X be a measure space. For discrete probability distributions p and q over the same domain X, it is defined as:

$$D_B(p,q) = -log(BC(p,q)) \qquad \text{(II.4)}$$

where

$$BC(p,q) = \sum_{x \in X} \sqrt{p(x)q(x)} \qquad \text{(II.5)}$$

is Bhattacharyya coefficient($0 \leq BC \leq 1$).

We will consider the characteristic vectors of the protein feature sequences and calculate their distances according to the Bhattacharyya distance. Advantages of using the Bhattacharyya distance are that:

1. It is computationally very simple;

2. It provides a "smoothed" distance between the two classes in study, which is more appropriate since we do not believe our data to be truly normally distributed.

By arranging all these values into a matrix, a pair-wise distance matrix is derived. This distance matrix contains the similarity information on the n protein primary structures. Lastly, this pair-wise distance matrix may be input to the Neighbour program(choosing the UPGMA method)in PHYLIP package[34] for a phylogenetic tree.

III. Experiments and Results

In this section, we will apply our method to real data to see how much phylogenetic information the feature sequences of proteins can extract. Generally, an independent method can be developed to evaluate the accuracy of a phylogenetic tree. Or the validity of a phylogenetic tree can be tested by comparing it with authoritative ones. Here, we adopt the latter one to test the validity of our phylogenetic trees.

A. Experiment No.1: Phylogenetic Analysis of Transferrins

In this experiment, we choose transferrin sequences from 24 vertebrates as a dataset[35]. Taxonomic information and accession numbers are provided in Table 1.

The feature sequences for the transferrin sequences are gained according to the mentioned rule in the second section. The evolutionary tree is generated by using the Neighbor joining(UPGMA) method in the PHYLIP package[34] . After discussing the value of k, we prefer $k = 6$ giving the best phylogeny. The result is shown in Fig.1. To indicate that the validity of our evolutionary trees, we show the result of Dai et al.[36]. Its result is shown in Fig.2. To compare our method with alignment method, we construct the evolutionary tree by ClustalW method. ClustalW, is a multiple sequence alignment program. The result is shown in Fig.3.

Table 1: Transferrin sequences, sources, and accession numbers.

Sequence Name	Species	Accession No.
Human TF	*Homo sapien*	S95936
Rabbit TF	*Oryctolagus coniculus*	X58533
Rat TF	*Rattus norvegicus*	D38380
Cow TF	*Bos Taurus*	U02564
Buffalo LF	*Bubalus arnee*	AJ005203
Cow LF	*Bos Taurus*	X57084
Goat LF	*Capra hircus*	X78902
Camel LF	*Camelus dromedaries*	AJ131674
Pig LF	*Sus scrofa*	M92089
Human LF	*H.sapiens*	NM_002343
Mouse LF	*Mus musculus*	NM_008522
Possum TF	*Trichosurus vulpecula*	AF092510
Frog TF	*Xenopus laevis*	X54530
Japanese flounder TF	*Paralichthys olivaceus*	D88801
Atlantic salmon TF	*Salmo salar*	L20313
Brown trout TF	*Salmo trutta*	D89091
Lake trout TF	*Salvelinus namaycush*	D89090
Brook trout TF	*Salvelinus fontinalis*	D89089
Japanese char TF	*Salvelinus pluvius*	D89088
Chinook salmon TF	*Oncorhynchus tshawytscha*	AH008271
Coho salmon TF	*Oncorhynchus hisutch*	D89084
Sockeye salmon TF	*Oncorhynchus nerka*	D89085
Rainbow trout TF	*Oncorhynchus mykiss*	D89083
Amago salmon TF	*Oncorhynchus masou*	D89086

*NOTE-TF, Transferring; LF, Lactoferrin.

Among three trees, the tree in Figure 1 is the most consistent with the classical trees constructed by Ford[35]. In Figure 2, the Rat TF, Cow TF are separated from Human TF and Rabbit TF, and lactoferrin (LF) proteins are assigned into two branches. This is contradict with the publicized existing trees. While Fig.3 also shows the unreasonable results. This verifies the validity of our method.

Summing up, our method gives a more intuitively acceptable arrangement, compared with the method of Dai et al. and the alignment-based method.

In addition, the whole process does not relate to complex algorithm and operation. Here, we compare the speed of our method with other methods by comparing their time complexity. In Table 2, we list the approximate estimation of time complexity of other algorithms. Table 2 shows that the time complexity of our model is favorable by comparing with that of the existing methods which solve the similar problem.

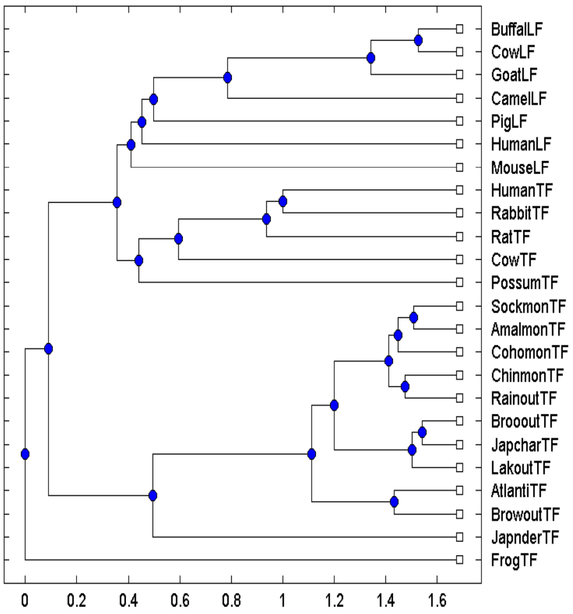

Fig. 1. Phylogenetic tree constructed by our method.

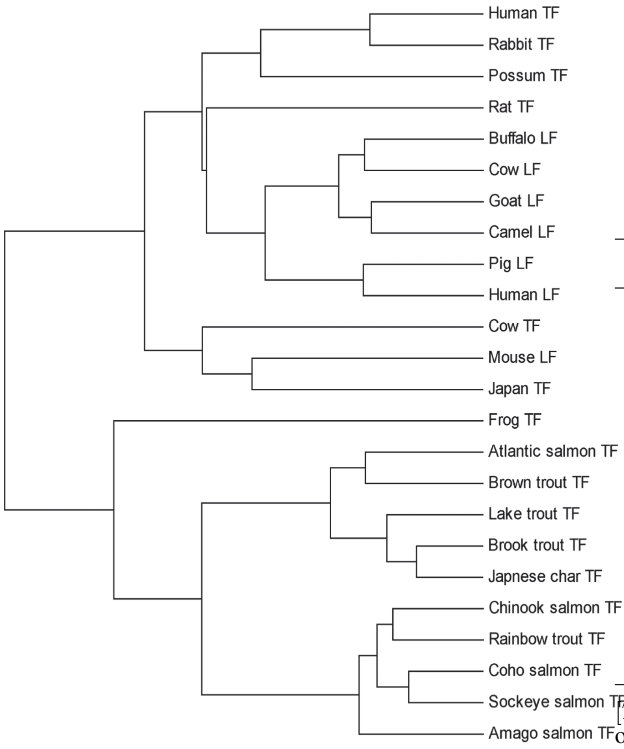

Fig. 2. The phylogenetic tree based on the distance of structural characteristic vector in Dai et al.

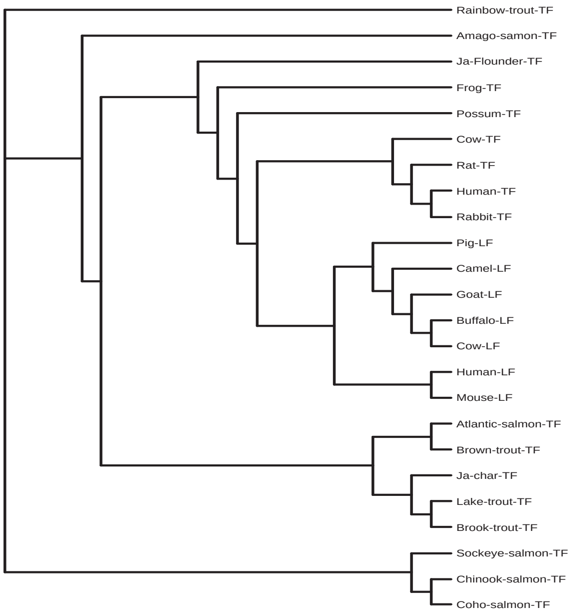

Fig. 3. Phylogenetic tree constructed by ClustalW.

Table 2 Comparison of time complexity of our method with other methods

References	Methods	The time complexity
Our method	Distance-based modeling	$O(n_1 + n_2)$
Shapiro and Zhang [37]	Tree comparison	$O([T_1][T_2])$
Corpet and Michot [38]	RNAlign program	$O(n_1^3 n_2^2)$
Bafna et al. [39]	Dynamic programming algorithms	$O(n_1^2 n_2^2)$
Dulucq and Tichit [40]	Tree edit algorithm	$O([T_1^{3/2}][T_2^{3/2}])$
Hofacker et al. [41]	Alignment of RNA base	$O(n_1^2 n_2^2)$
Yao et al. [42]	Leading eigenvalues of E matrix	$O(n_1^3 + n_2^3)$
Yao et al. [43]	Leading eigenvalues of D/D matrix	$O(n_1^3 + n_2^3)$
Zhu et al.[44], Bai and Wang [45]	Leading eigenvalues of L/L matrix	$O(n_1^2 + n_2^2)$

$[T_i]$ is the number of nodes in the tree T_i; F_i is the number of nodes in the forest F_i and $deg(F_i)$ is the degree of F_i; n_i denotes the size of ith sequence.

B. Experiment No.2: Phylogenetic Analysis of Coronavirus Spike Proteins

In order to further verify the validity of our method, in this experiment, we turn to make phylogenetic analysis of the 26 spike protein sequences from coronavirus. Taxonomic information and accession numbers are provided in Table 3.

Table 3 Coronavirus spike proteins sequences, sources, and accession numbers.

Sequence Name	Species	Accession
TGEV	*Transmissible gastroenteritis virus*	NP_058424
PEDV	*Porcine epidemic diarrhea virus*	NP_598310
HCoV-OC43	*Human coronavirus OC43*	NP_937950
BCoVM	*Bovine coronavirus strain Mebus*	AAA66399
BCoVL	*Bovine coronavirus isolate BCoV-LUN*	AAL57308
BCoVQ	*Bovine coronavirus strain Quebec*	AAL40400
BCOV	*Bovine coronavirus*	NP_150077
MHVM	*Mouse hepatitis virus strain ML-10*	AAF69344
MHVP	*Mouse hepatitis virus strain Penn 97-1*	AAF69334
MHVJHM	*Murine hepatitis virus strain JHM*	YP_209233
MHVA	*Mouse hepatitis virus strain MHV-A59C12 mutant*	AAB86819
IBVBJ	*Avain infectious bronchitis virus isolate BJ*	AAP92675
IBVC	*Avain infectious bronchitis virus strain Ca199*	AAS00080
IBV	*Avain infectious bronchitis virus*	NP_040831
GD03T0013	*SARS coronavirus GD03T0013*	AAS10463
PC4-127	*SARS coronavirus PC4-127*	AAU93318
PC4-137	*SARS coronavirus PC4-137*	AAV49720
Civet007	*SARS coronavirus civet007*	AAU04646
A022	*SARS coronavirus A022*	AAV91631
GD01	*SARS coronavirus GD01*	AAP51227
GZ02	*SARS coronavirus GZ02*	AAS00003
CUHK-W1	*SARS coronavirus CUHK-W1*	AAP13567
TOR2	*SARS coronavirus TOR2*	AAP41037
Urbani	*SARS coronavirus Urbani*	AAP13441
Frankfurt1	*SARS coronavirus Frankfurt1*	AAP33697
Sino1-11	*SARS coronavirus Sino1-11*	AAR23250

The phylogenetic tree for the 26 spike proteins from coronavirus is constructed by our method, which is presented in Fig.4. From Fig.4 we can see that the SARS-CoVs appear to cluster together and form a separate branch, which can be easily distinguished from other three groups of coronaviruses.

In order to compare our method with alignment-based method, we also construct the phylogenetic tree by ClustalW method. The result is shown in Fig.5. Compared with the two results, we can see that the phylogenetic tree constructed by our method is more consistent with the known fact of evolution[46, 47].

IV. Conclusions and discussion

With the development of the technology, more and more biological sequences have been collected for analysis. In the present study, we introduce the phylogenetic analysis of protein sequences based on the characteristic vectors of protein feature sequences and the Bhattacharyya distance. In this paper, we integrate

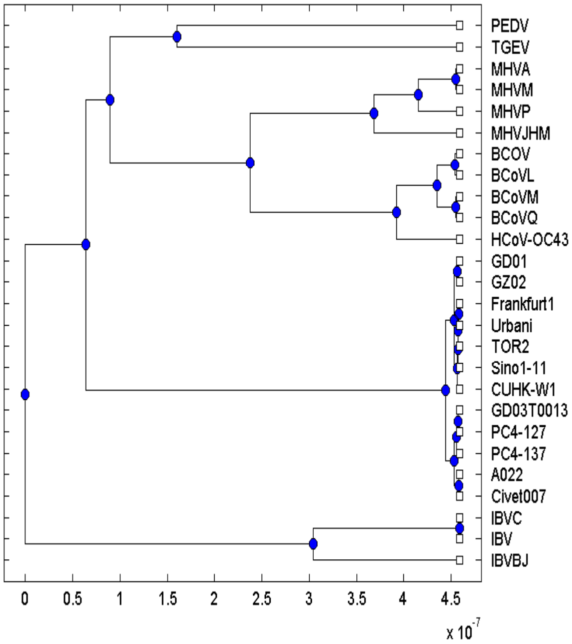

Fig. 4. Phylogenetic tree constructed by our method.

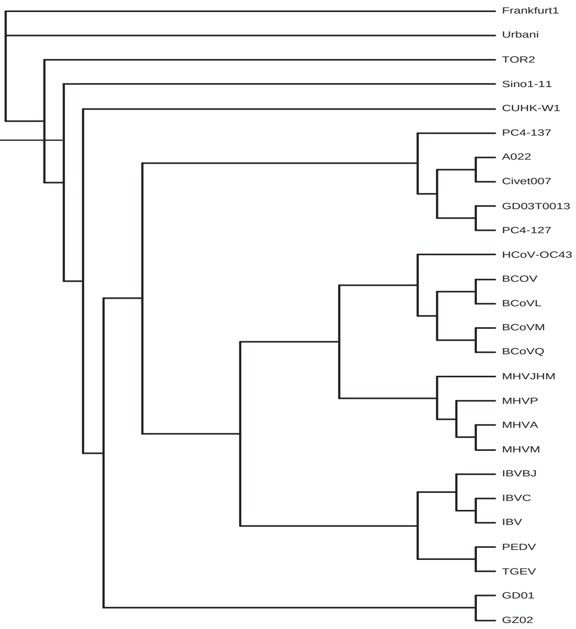

Fig. 5. Phylogenetic tree constructed by ClustalW.

the physicochemical properties of amino acids into the Bhattacharyya distance to phylogenetic analysis. The Bhattacharyya distance is a theoretical distance measure between two probability distributions. It also has the desirable properties of being computationally simple. Our examples have indicated that the introduction of the protein feature sequences into evolution analysis is successful.

In a word, it is a novel alignment-free method that yields results reasonably and rapidly. Our method is not necessarily an improvement as compared to some existing methods, but an alternative for phylogenetic analysis of protein sequences. The new method does not require sequence alignment and the construction of tree models. The shortage of this method is that some information may be lost in the protein feature sequences. However, our tests have proven that our method can be served as an alternative tool among other alignment-based and alignment-free methods for phylogenetic analysis of protein sequences.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China with the Grant No.10871219.

REFERENCES

[1] Y. Lin, S. Fang, J. Thorne, "A tabu search algorithm for maximum parsimony phylogeny inference", *Eur. J. Oper. Res.* 176, 2007, pp. 1908–1917.

[2] F. Ren, H. Tanaka, Z. Yang, "A likelihood look at the supermatrix-supertree controversy", *Gene.* 441, 2009, pp. 119–125.

[3] A. Som, "ML or NJ-MCL? A comparison between two robust phylogenetic methods", *Comput. Biol. Chem.* 33, 2009, pp. 373–378.

[4] M. B. Elliott, D. M. Irwin, E. P. Diamandis, "In silico identification and bayesian phylogenetic analysis of multiple new mammalian kallikrein gene families", *Genomics.* 88, 2006, pp. 591–599.

[5] E. Jako, E. Ari, P. Ittzes, A. Horvath, J. Podani, "BOOL-AN: A method for comparative sequence analysis and phylogenetic reconstruction", *Mol. Phy. Evol.* 52, 2009, pp. 887–897.

[6] S. Zhang, T. Wang, "Feature analysis of protein structure by using discrete Fourier transform and continuous wavelet transform", *J Math Chem.* 46, 2009, pp. 562–568.

[7] S. Zhang, T. Wang, "A complexity-based method to compare RNA secondary structures and its application", *Journal of Biomolecular Structure and Dynamics*, 28(2), 2010, pp. 247–258.

[8] L. Yang, G. Chang, X. Zhang, T. Wang, "Use of the Burrows-Wheeler similarity distribution to the comparison of the proteins", *Amino Acids*, 39(3), 2010, pp. 887–898,

[9] Z. Cao, B. Liao, R. Li, "A group of 3D graphical representation of DNA sequences based on dual nucleotides", *Int. J. Quantum. Chem.* 108, 2008, pp. 1485–1490.

[10] Z. Liu, B. Liao, W. Zhu, "A new method to analyze the similarity based on dual nucleotides of the DNA sequence", *MATCH Commun. Math. Comput. Chem.* 61, 2009, pp. 541–552.

[11] W. Zhu, B. Liao, R. Li, "A novel method for constructing phylogenetic tree based on a dissimilarity matrix", *MATCH Commun. Math. Comput. Chem.* 63, 2010, pp. 483–492.

[12] B. Liao, L. Liao, G. Yue, R. Wu, W. Zhu, "A vertical and horizontal method for constructing phylogenetic tree", *MATCH Commun. Math. Comput. Chem.* 63, 2010, pp. 691–700.

[13] S. Zhang, L. Yang, T. Wang, "Use of information discrepancy measure to compare protein secondary structures", *J. Mol. Struct: THEOCHEM.* 909, 2009, pp. 102–106.

[14] S. Zhang, T. Wang. "Phylogenetic Analysis of Protein Sequences Based on Conditional LZ Complexity". *MATCH Commun. Math. Comput. Chem.* 63, 2010, pp. 701–716.

[15] R. I. Mubark, H. A. Keshk, M. I. Eladawy, "Different Species Classifier and Hemoglobin Structure Predictor based on DNA Sequences", *International Journal of Biology and Biomedical Engineering*, 2(2), 2008, pp. 49–58.

[16] R. I. Mubark, H. A. Keshk, M. I. Eladawy, "Different Species Classifier and Hemoglobin Structure Predictor based on DNA Sequences", *International Journal of Biology and Biomedical Engineering*, 2(2), 2008, pp. 98–107.

[17] C. Huang, C. Lin, H. Jan, "System Identification and Control Using DNA Computing Algorithms", *International Journal of Biology and Biomedical Engineering*, 4(2), 2008, pp. 108–117.

[18] R. I. Mubark, H. A. Keshk, M. I. Eladawy, "Different Species and Proteins Classifiers and Protein's Structure Predictors Systems", *International Journal of Biology and Biomedical Engineering*, 4(2), 2008, pp. 119–128.

[19] R. Ivancsy, I. Vajk, "PD-Tree: A new approach to subtree discovery", *WSEAS transactions on infor-*

mation science and applications, 11(2), 2005, pp. 1772–1779.

[20] F. Bai, T. Wang, “The construction of phylogenetic tree by Graphic Representation of DNA Sequences”, *Proceedings of the 5th WSEAS Int. Conf. on simulation, modeling and optimization*, Corfu, Greece, August 17-19, 2005, pp. 463–467.

[21] K. Lin, C. Y. Lin, C.D. Huang, etc., “Improving Prediction Accuracy for Protein Structure Classification by Neural Network Using Feature Combination”, *Proceedings of the 5th WSEAS Int. Conf. on applied informatics and communications*, September 15-17, 2005, pp. 313–318.

[22] N. Todorova, A. Hung, I. Yarovsky, “Application of Computational Modelling to Protein Folding and Aggregation Studies”, *Proceedings of the 10th WSEAS International Conference on mathematics and computers in biology and chemistry*, 2009, pp. 130–135.

[23] T. F. Gharib, A. Salah, A. M. Salem, “PSISA: An Algorithm for Indexing and Searching Protein Structure using Suffix Arrays”, *Proceedings of the 12th WSEAS International Conference on computers*, Heraklion, Greece, July 23-25, 2008, pp. 775–780.

[24] H. Zhang, Y. Zhong, B. Hao, X. Gu, “A simple method for phylogenomic inference using the information of gene content of genomes”, *Gene.* 441, 2009, pp. 163–168.

[25] Z. Yu, V. Anh, L. Zhou, “Fractal and dynamical language methods to construct phylogenetic tree based on protein sequences from complete genomes”, *Advances in Natural Computation. PT3, Proceedings.* 3612, 2005, pp. 337–347.

[26] D. R. Bastola, H. H. Otu, S. E. Doukas, K. Sayood, S. H. Hinrichs, P. C. Iwen, “Utilization of the relative complexity measure to construct a phylogenetic tree for fungi”, *Mycol. Res.* 108, 2004, pp. 117–125.

[27] N. Liu, T. Wang, “Protein-based phylogenetic analysis by using hydropathy profile of amino acids”, *FEBS Lett.* 580, 2006, pp. 5321–5327.

[28] T. Kailath, “The Divergence and the Bhattacharyya distance measures in signal selection”. *IEEE Trans. Commun. Technol.* C-15, 1967, pp. 52–60.

[29] A. Bhattacharyya, “On a measure of divergence between two statistical populations defined by their probability distributions”. *Bull. Calcutta. Math. Soc.* 49, 1943, pp. 214–224.

[30] Z. Yu, V. Anh, K. Lau, “Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses”, *J. Theor. Biol.* 226, 2004, pp. 341–348.

[31] S. Karlin, M. Ladunga, “Comparisons of eukaryotic genomic sequences”. *Proc. Natl. Acad. Sci.* 91, 1994, pp. 12832–12836.

[32] L. Gao, J. Qi, B. L. Hao, “Simple markov subtraction essentially improves prokaryote phylogeny”. *AAPPS B.* June, 2006, pp. 3–7.

[33] J. Qi, B. Wang, B. L. Hao, “Whole proteome prokaryote phylogeny without sequence alignment: A K-String composition approach”. *J. Mol. Biol.* 58, 2004, pp. 1–11.

[34] J. Felsenstein, “PHYLIP-phylogeny inference package (version 3.2)”, *Cladistics* 5, 1989, pp. 164–166.

[35] M. Ford, “Molecular evolution of transferrin: Evidence for positive selection in salmonids”, *Mol. Biol. Evol.* 18, 2001, pp. 639–647.

[36] Q. Dai, X. Liu, T. Wang, “Analysis of protein sequences and their secondary structures based on transition matrices”, *J. Mol. Struct: THEOCHEM.* 803, 2007, pp. 115–122.

[37] B. Shapiro, K. Zhang, “Comparing multiple RNA secondary structures using tree comparisons”, *Comput. Appl. Biosci.* 6, 1990, pp. 309–318.

[38] F. Corpet, B. Michot, “RNAlign program: alignment of RNA sequences using both primary and secondary structures”, *Comput. Appl. Biosci.* 10, 1995, pp. 389–399.

[39] V. Bafna, S. Muthukrishnan, R. Ravi, “Computer similarity between RNA strings”. *Proceedings of the 6th Symposium on Combinatorial Pattern Matching*, CPM-95, 1995, pp. 1–16.

[40] S. Dulucq, L. Tichit, “RNA secondary structure comparison: exact analysis of the Zhang-Shasha tree edit algorithm”. *Theor. Comput. Sci.* 306, 2003, pp. 471–484.

[41] I. L. Hofacker, S. H. F. Bernhart, P. F. Stadler, “Alignment of RNA base pairing probability matrices”. *Bioinformatics* 20, 2004, pp. 2222–2227.

[42] Y. H. Yao, X. Y. Nan, T. M. Wang, “A class of 2D graphical representations of RNA secondary structures and the analysis of similarity based on them”. *J. Comput. Chem.* 26, 2005, pp. 1339–1346.

[43] Y. H. Yao, X. Y. Nan, T. M. Wang, “A 2D graphical representation of RNA secondary structures and the analysis of similarity/dissimilarity based on it”. *J. Mol. Struc. Theochem.* 755, 2005, pp. 131–136.

[44] W. Zhu, B. Liao, K. Q. Ding, “A condensed 3D graphical representation of RNA secondary

structures". *J. Mol. Struc. Theochem.* 757, 2005, pp. 193–198.

[45] F. Bai, T. M. Wang, "On graphical and numerical representation of protein sequences". *J. Biomol. Struc. Dyn.* 23, 2006, pp. 537–545.

[46] W. X. Zheng, L. L. Chen, H. Y. Ou, F. Gao, C. T. Zhang, "Coronavirus phylogeny based on a geometric approach". *Mol Phylogenet Evol.* 36, 2005, pp. 224–232.

[47] H. D. Song, C. C. Tu, G. W. Zhang, et al. "Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human". *PNAS.* 102, 2005, pp. 2430–2435.

Shengli Zhang is with the School of Mathematical Sciences, Dalian University of Technology, P.R.China. No.2 Linggong Road, Ganjingzi District, Dalian, 116024, P.R.China. (Phone: +86-411-84749735. Fax: +86-411-84708354 E-mail: shengli0201@163.com).

Tianming Wang is with the School of Mathematical Sciences, Dalian University of Technology, P.R.China. No.2 Linggong Road, Ganjingzi District, Dalian, 116024, P.R.China. (Phone: +86-411-84749735. Fax: +86-411-84708354 E-mail: wangtm@dlut.edu.cn).