MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **63** (2010) 701-716

ISSN 0340 - 6253

Phylogenetic Analysis of Protein Sequences Based on Conditional LZ Complexity

Shengli Zhang,[a] **Tianming Wang**

[a] School of Mathematical Sciences, Dalian University of Technology
Dalian 116024, P.R.China
e-mail: shengli0201@163.com

(Received November 16, 2009)

Abstract

Up to now, various approaches for phylogenetic analysis have been developed. Almost all of them put stress on analyzing nucleic acid sequences or protein primary structures. In this paper, we take the physicochemical properties of amino acids into account and introduce the protein feature sequences into phylogenetic analysis by using conditional LZ complexity. We find that this method is effectual and feasible.

INTRODUCTION

Protein is composed of amino acids, and it is the amino acid sequence that determines the chemical structure of protein. The protein sequence analysis can be used to find homologous proteins,classify protein families,and construct evolutionary tree [1–3]. Phylogenetics is the study of the evolutionary history among organisms. Moreover, it can provide information for function prediction and pharmaceutical researchers may use phylogenetic methods to determine species, thus perhaps sharing their medicinal qualities [4].

Some researchers explore many methods for phylogenetic analysis, for instance, distance methods, maximal parsimony methods, maximum likelihood methods and Bayesian methods [5–9], each of which has its own range of applicability. Biologists and researchers are always trying to develop efficient methods for complex phylogenetic analysis [10–17].

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 [18]. Yu et al. used the multiplicative model to analyze character string frequencies and derive phylogenies, where each protein was represented by a composition vector [19]. 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 [20]. 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 [2]. 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. LZ algorithm is a widely used alignment-free algorithm and we can calculate the complexity of a single sequence according to the LZ algorithm. To depict the complexity relationship between two sequences, in this paper, we use conditional LZ algorithm to analyze the phylogenethy of protein sequences.

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 [21]. 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 letters 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.

SEQUENCE CONDITIONAL LZ COMPLEXITY AND DISTANCE METRIC

LZ algorithm was developed to analyze the complexity of linear sequences by Lempel and Ziv in 1976 [22]. In recent years, some authors applied the algorithm construct phylogenic tree. For instance, Otu et al. applied LZ algorithm to phylogenic analysis and had successfully constructed phylogenic trees for real and simulated DNA data sets [23]. Liu and Wang take the physicochemical properties of amino acids into account, and used

LZ algorithm to construct phylogenic trees [2]. Li et al. also used the conditional LZ complexity to analyze DNA sequences and to reconstruct phylogenetic tree [24]. Then, we will give some basic definitions.

LZ COMPLEXITY

Let the digital sequence $S = s_1 s_2 \cdots s_n, l(S) = n$ represent the length of S, the subsequence $s_i s_{i+1} \cdots s_j$ of S be denoted as $S(i, j)$. Note that $S(i, j) = \emptyset$, for $i > j$. The set that contains all subsequence $S(i, j)$ is called the vocabulary $v(S)$ of S. According to the computable modeling proposed by Lempel and Ziv [22], the sequence S can partition into some subsequences that arrange one after another. Denote this partition as follows:

$$H_{LZ}(S) = S(h_0 + 1, h_1)S(h_1 + 1, h_2) \cdots S(h_k + 1, h_{k+1}) \cdots S(h_{m-1} + 1, h_m)$$

$H_{LZ}(S)$ satisfies the following three properties:

(1) $h_0 = 0, h_1 = 1$;

(2) $\forall 1 \leq k \leq (m-2)$,
$S(h_k + 1, h_{k+1} - 1) \in v(S(1, h_{k+1} - 2))$,
$S(h_k + 1, h_{k+1}) \notin v(S(1, h_{k+1} - 1))$;

(3) $h_m = l(S)$,
$S(h_{m-1} + 1, h_m - 2) \in v(S(1, h_m - 2))$.

Lempel and Ziv proved the exclusive partition about H_{LZ} and defined the complexity $c_{LZ}(S)$ of S as the number of subsequence in $H_{LZ}(S)$, namely $c_{LZ}(S) = m$. The flow diagram [25] for the algorithm to calculate c_{LZ} is shown by Fig.1. The time complexity of this algorithm is $O(l(S)^2)$.

CONDITIONAL LZ COMPLEXITY

LZ algorithm is valid to describe the complexity of single sequence, but it can not describe the complex relationships between two sequences. To depict sequences more clearly, Li et al. [24] proposed the definition of conditional LZ complexity.

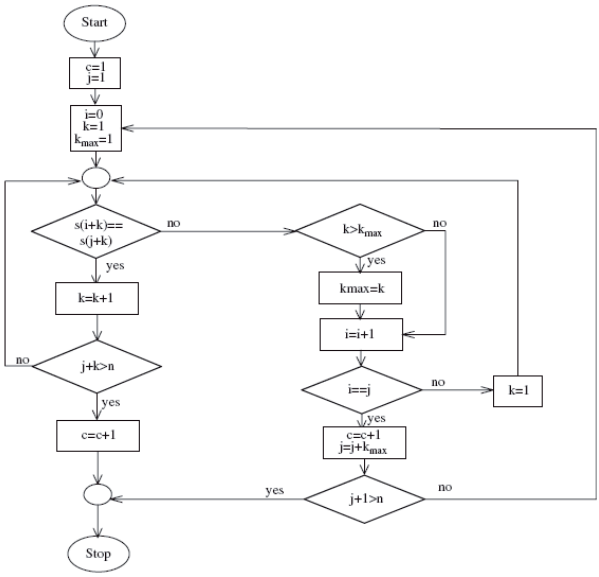

Figure 1: Flow diagram for the algorithm to calculate c_{LZ}.

Given two digital sequences S, T and sequence T as a conditional sequence, according to the theory of Lempel and Ziv about sequence partition, we can also partition sequence S into the subsequences one after another, called it conditional partition of S corresponding to conditional sequence T. Denote it as follows:

$$H_{LZ}(S \mid T) = S(h_0 + 1, h_1)S(h_1 + 1, h_2) \cdots S(h_k + 1, h_{k+1}) \cdots S(h_{m'-1} + 1, h_{m'})$$

$H_{LZ}(S \mid T)$ satisfies the following three properties:

(1) $h_0 = 0$;

(2) $\forall 1 \leq k \leq (m^{'} - 2)$,
$S(h_k + 1, h_{k+1} - 1) \in v(TS(1, l(T) + h_{k+1} - 2))$,
$S(h_k + 1, h_{k+1}) \notin v(TS(1, l(T) + h_{k+1} - 1))$;

(3) $h_{m'} = l(S)$,
$S(h_{m'-1} + 1, h_{m'} - 2) \in v(TS(1, l(T) + h_{m'} - 2))$.

The complexity $c_{LZ}(S \mid T)$ of S corresponding to conditional sequence T is the number of subsequence in $H_{LZ}(S \mid T)$, namely $c_{LZ}(S \mid T) = m^{'}$. This conditional partition is

also exclusive. To compute the conditional LZ complexity of S, we only need add the conditional sequence T as the prefix to the pattern search area. The time complexity of this algorithm is $O(l(S) \times [l(T) + l(S)])$.

DISTANCE METRIC

According to the conditional LZ complexity, Li et al. defined the distance metric between two digital sequences. Given non-null sequences S and T, their complexity distance is

$$D(S,T) = max\{c_{LZ}(S \mid T), c_{LZ}(T \mid S)\}$$

It had been proved that the complexity distance satisfies the following four properties about distance metric based on it can add constant 1:

(1) $D(S,T) > 0, \forall S \neq T$;

(2) $D(S,T) = 0, \forall S = T$;

(3) $D(S,T) = D(T,S), \forall S,T$;

(4) $D(S,T) \leq D(S,R) + D(R,T), \forall R,S,T$.

In this paper, to eliminate the computation difference of distance metric generated by the length of data, we use the normalized distance $D^{'}(S,T)$ as the last distance metric:

$$D^{'}(S,T) = D(S,T)/(l(S) + l(T))$$

namely,

$$D^{'}(S,T) = \begin{cases} max\{c_{LZ}(S \mid T), c_{LZ}(T \mid S)\}/(l(S) + l(T)), & S \neq T \\ 0 & S = T. \end{cases}$$

We will consider the protein feature sequences and calculate their distances according to the above equation. 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 [26] for a phylogenetic tree.

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.

EXPERIMENT NO.1: PHYLOGENETIC ANALYSIS OF TRANSFERRINS

In the first experiment, we choose transferrin sequences from 24 vertebrates as a dataset [27]. Taxonomic information and accession numbers are provided in Table 1.

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.

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 [26] . The result is shown in Fig.2. To indicate that the validity of our evolutionary trees,we show the result of Dai et al. [28]. Its result is shown in Fig.3. To compare conditional LZ method with alignment method, we constructed the evolutionary tree by ClustalW method. ClustalW, is a multiple sequence alignment program. The result is shown in Fig.4.

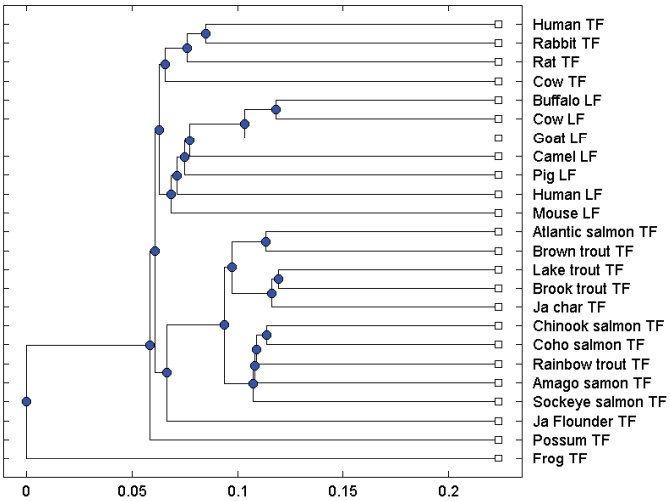

Figure 2: The phylogenetic tree constructed by our method

Compared with the result in Fig.2 and Fig.3, we find ours is better:

1. From Fig.2 we can observe that all the proteins that belong to transferrin(TF) proteins and lactoferrin(LF) proteins have been separated well and grouped into respective taxonomic classes accurately. The tree in Fig.2 is the most consistent with the trees constructed by Ford [27], which is the most classical result in the publicized existing trees. This verifies the validity of our method.

2. In Fig.2, the Human TF, Rabbit TF, Rat TF and Cow TF are clustered into the same branch while in Fig.3, the Rat TF, Cow TF are separated from Human TF and Rabbit TF, this contradicts the classical result.

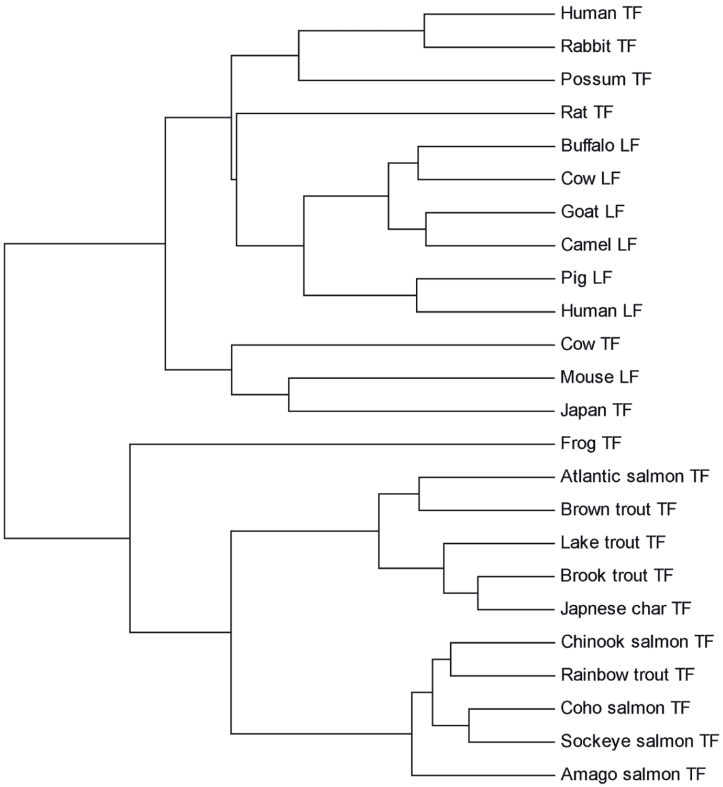

Figure 3: The phylogenetic tree based on the distance of structural characteristic vector in Dai et al.(20)

3. Fig.3 shows and lactoferrin(LF) proteins are assigned into two different branches. This contradicts the traditional opinion and the advantage of our method is more obvious.

Compared with the result in Fig.2 and Fig.4, we find ours is also better:

1. The transferrin(TF) proteins and lactoferrin(LF) proteins are clustered into their corresponding branches in Fig.2, while they are mixed together in Fig.4 and they are far with each other. This contradicts the traditional opinion.

2. In respect to the transferrin(TF) proteins, our result in Fig.2 is better than Fig.4 in general. That shows our result is more closed to classical results.

3. In respect to the lactoferrin(LF) proteins, the two methods are almost the same.

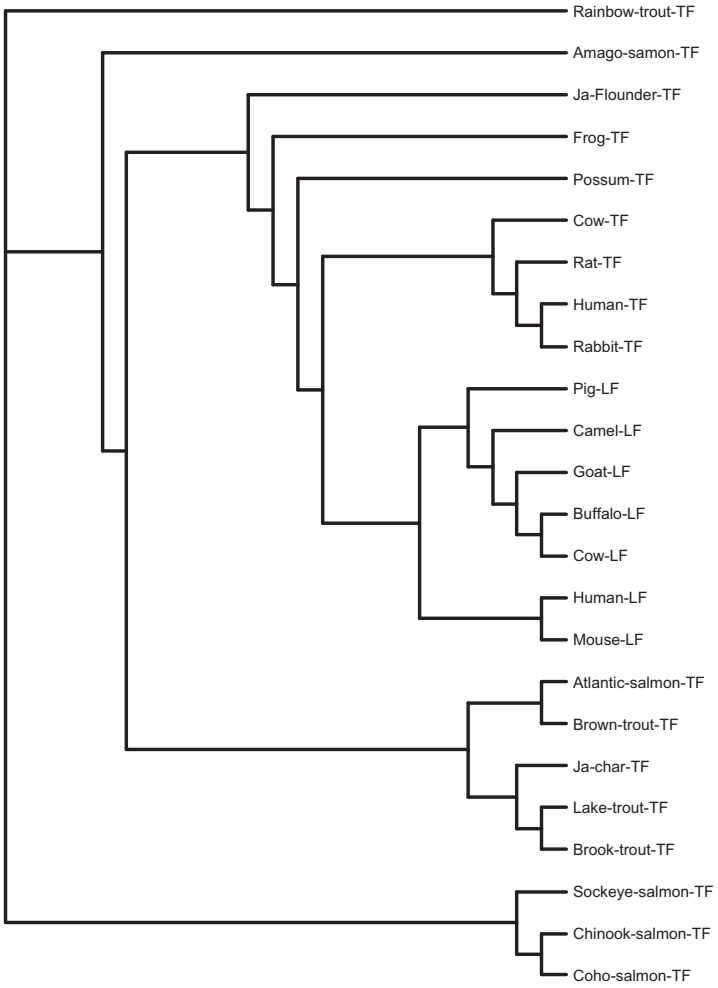

Figure 4: The phylogenetic tree constructed by ClustalW

Summing up, our method has significant advantage, compared with the method of Dai et al. and the alignment-based method.

EXPERIMENT NO.2: PHYLOGENETIC ANALYSIS OF ND5(NADH DEHYDROGENASE SUBUNIT 5)PROTEINS

In order to further verify the validity of our method, in this experiment, we turn to make phylogenetic analysis of sequences belonging to nine ND5(NADH dehydrogenase subunit 5)proteins: human(*Homo sapiens*, AP_000649), gorilla(*Gorilla gorilla*, NP_008222), common chimpanzee(*Pan troglodytes*, NP_008196), pigmy chimpanzee(*Pan paniscus*, NP_008209), fin whale(*Balenoptera physalus*, NP_006899), blue whale(*Balenoptera musculus*, NP_007066), rat(*Rattus norvegicus*, AP_004902), mouse(*Mus musculus*, NP_904338), and opossum(*Didelphis virginiana*, NP_007105).

The phylogenetic tree for ND5 proteins is constructed by our method, which is presented in Fig.5. In order to compare conditional LZ method with alignment method, we also constructed the evolutionary tree by ClustalW method. The result is shown in Fig.6.

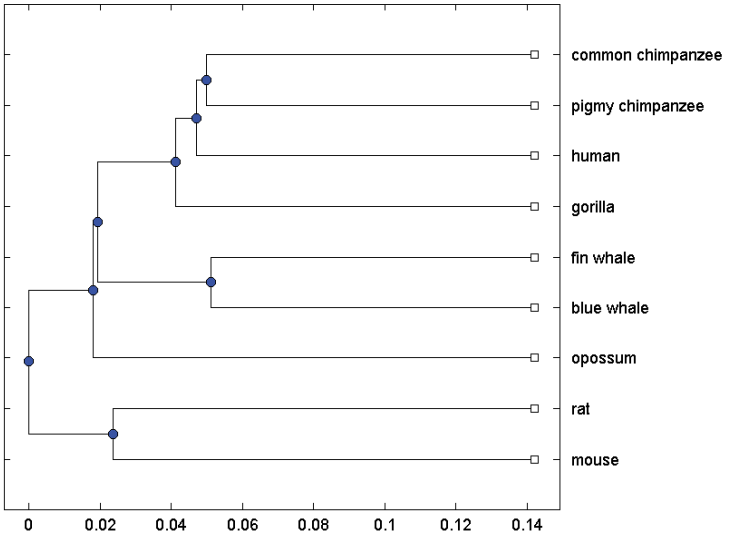

Figure 5: The phylogenetic tree for the ND5 proteins based on our method

Compared with the result in Fig.5 and Fig.6, we can see that the phylogenetic tree constructed by our method is more consistent with the known fact of evolution [23, 29, 30]:

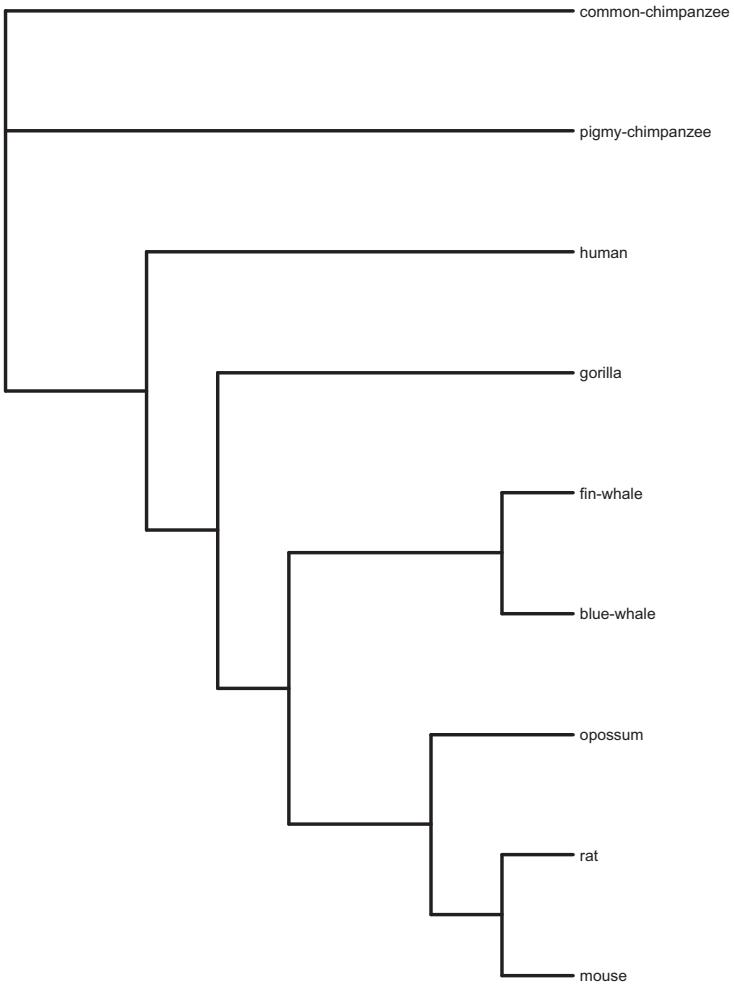

Figure 6: The phylogenetic tree constructed by ClustalW

1. From Fig.5 we can see that the ND5 proteins of human, gorilla, common chimpanzee, pigmy chimpanzee are more similar with each other, and they are clustered into the different branches in Fig.6. This contradicts the traditional opinion.

2. Furthermore, the ND5 proteins of human is more similar to common chimpanzee and pigmy chimpanzee than gorilla in Fig.5. This is consistent with the known fact of evolution.

3. The fin whale and blue whale, rat and mouse are also similar, respectively. The two methods are almost the same.

CONCLUSIONS AND DISCUSSIONS

With the development of the technology, more and more biological sequences have been collected for analysis. Conditional LZ algorithm has been introduced into protein feature sequences studies. The main advantage is that this algorithm can extract repeated patterns from biological sequences. Therefore, when two sequences are compared, the subsequence that they share can be detected. In this paper, we integrate the physicochemical properties of amino acids into conditional LZ algorithm to phylogenetic analysis, because conditional LZ algorithm can extract more efficient information between two sequences than LZ algorithm. Our examples have indicated that the introduction of the protein feature sequences into evolution analysis is successful.

The shortage of this method is that some information may be lost when protein primary structures are converted to protein feature sequences. However, our tests have proven that our method can extract phylogenetic information from proteins.

Acknowledgements: The authors thank the anonymous referees for many valuable suggestions that have improved this manuscript. We appreciate the financial support of this work that was provided by the National Natural Science Foundation of China with the grant No.10871219.

References

[1] Q. Xu, A. A. Canutescu, G. Wang, M. Shapovalov, Z. Obradovic, R. L. Dunbrack, Statistical analysis of interface similarity in crystals of homologous proteins, *J. Mol. Biol.* **381** (2008) 487–507.

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

[3] C. Jia, T. Liu, X. Zhang, H. Fu, Q. Yang, Alignment-free comparison of protein sequences based on reduced amino acid alphabets, *J. Biomol. Str. Dyn.* **6** (2009) 26–32.

[4] K. Komatsu, S. Zhu, H. Fushimi, T. K. Qui, S. Cai, S. Kadota, Phylogenetic analysis based on 18S rRNA gene and matK gene sequences of Panax vietnamensis and five related species, *Planta Med.* **67** (2001) 461–465.

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

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

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

[8] 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) 591–599.

[9] 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) 887–897.

[10] B. Liao, Y. Liu, R. Li, W. Zhu, Coronavirus phylogeny based on triplets of nucleic acids bases, *Chem. Phys. Lett.* **421** (2006) 313–318.

[11] B. Liao, X. Xiang, W. Zhu, Coronavirus phylogeny based on 2D graphical representation of DNA sequence, *J. Comput. Chem.* **27** (2006) 1196–1202.

[12] B. Liao, X. Shan, W. Zhu, R. Li, Phylogenetic tree construction based on 2D graphical representation, *Chem. Phys. Lett.* **422** (2006) 282–288.

[13] 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) 1485–1490.

[14] 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) 541–552.

[15] 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) 483–492.

[16] 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).

[17] S. Zhang, L. Yang, T. Wang, Use of information discrepancy measure to compare protein secondary structures, *J. Mol. Struct. (Theochem)* **909** (2009) 102–106.

[18] 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) 163–168.

[19] 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) 337–347.

[20] 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) 117–125.

[21] 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) 341–348.

[22] A. Lempel, J. Ziv, On the complexity of finite sequences, *IEEE Trans. Inform. Theory* **22** (1976) 75–81.

[23] H. H. Otu, K. Sayood, A new sequence distance measure for phylogenetic tree construction, *Bioinformatics* **19** (2003) 2122–2130.

[24] B. Li, Y. Li, H. He, LZ complexity distance of DNA sequence and its application in phyligenetic tree reconstruction, *Geno. Prot. Bioinfo.* **3** (2005) 206-212.

[25] L. Liu, T. Wang, Comparison of TOPS strings based on LZ complexity, *J. Theor. Biol.* **251** (2008) 159–166.

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

[27] M. Ford, Molecular evolution of transferrin: Evidence for positive selection in salmonids, *Mol. Biol. Evol.* **18** (2001) 639–647.

[28] Q. Dai, X. Liu, T. Wang, Analysis of protein sequences and their secondary structures based on transition matrices, *J. Mol. Struct. (Theochem)* **803** (2007) 115–122.

[29] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, H. Zhang, An information-based sequence distance and its application to whole mitochondrial genome phylogeny, *Bioinformatics* **17** (2001) 149–154.

[30] V. Makarenkov, F. Lapointe, A weighted least-squares approach for inferring phylogenies from incomplete distance matrices, *Bioinformatics* **20** (2004) 2113–2121.