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 has 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 \(i\)th 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 sequence \(S=VFFPDETGTGSYHRWGSTQQCQVFEGLDEQQ\), its feature sequence is \(R(S)=000011222223030222222001201122\).

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.

\[\alpha(a_1a_2\ldots a_k) = \begin{cases} 
\frac{f(a_1a_2\ldots a_k) - \eta_0(a_1a_2\ldots a_k)}{\eta_0(a_1a_2\ldots a_k)}, & f^0(a_1a_2\ldots a_k) \neq 0; \\
0, & f^0(a_1a_2\ldots a_k) = 0.
\end{cases}\]

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 \((\alpha_1, \alpha_2, \ldots, \alpha_{4^k})\), where \(\alpha_i\) means the characteristic of the \(i\)th \(k\)-word. We can construct a \(k\)-word characteristic vector \(A_k = (\alpha_1^A, \alpha_2^A, \ldots, \alpha_{4^k}^A)\) for sequence \(A\) and likewise \(B_k = (\alpha_1^B, \alpha_2^B, \ldots, \alpha_{4^k}^B)\) 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:

$$D_B(p, q) = -\log(BC(p, q)) \quad (II.4)$$

where

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

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 pairwise distance matrix is derived. This distance matrix contains the similarity information on the $n$ protein primary structures. Lastly, the 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.

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

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

$T_i$ is the number of nodes in the tree $T_i$; $F_i$ is the number of nodes in the forest $F_i$; $deg(F_i)$ is the degree of $F_i$; $n_i$ denotes the size of $i$th 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>
<thead>
<tr>
<th>Sequence Name</th>
<th>Species</th>
<th>Accession No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGEV</td>
<td>Transmissible gastroenteritis virus</td>
<td>NP_058424</td>
</tr>
<tr>
<td>PEDV</td>
<td>Porcine epidemic diarrhea virus</td>
<td>NP_598310</td>
</tr>
<tr>
<td>HCoV-OC43</td>
<td>Human coronavirus OC43</td>
<td>NP_927950</td>
</tr>
<tr>
<td>BCoVL</td>
<td>Bovine coronavirus isolate BCoV</td>
<td>AAL57308</td>
</tr>
<tr>
<td>BCoVQ</td>
<td>Bovine coronavirus strain Quebec</td>
<td>AAL40040</td>
</tr>
<tr>
<td>BCOV</td>
<td>Bovine coronavirus</td>
<td>NP_150977</td>
</tr>
<tr>
<td>MHVVM</td>
<td>Mouse hepatitis virus strain ML-10</td>
<td>AAF69344</td>
</tr>
<tr>
<td>MHVVP</td>
<td>Mouse hepatitis virus strain Penn 97-1</td>
<td>AAF69334</td>
</tr>
<tr>
<td>MHVJHM</td>
<td>Murine hepatitis virus strain JHM</td>
<td>YP_209233</td>
</tr>
<tr>
<td>MHA</td>
<td>Mouse hepatitis virus strain MHV- A59C12 mutant</td>
<td>AAS68619</td>
</tr>
<tr>
<td>IBVBJ</td>
<td>Avian infectious bronchitis virus isolate BJ</td>
<td>AAPP2675</td>
</tr>
<tr>
<td>IBVC</td>
<td>Avian infectious bronchitis virus strain Ca199</td>
<td>AAS00080</td>
</tr>
<tr>
<td>IBV</td>
<td>Avian infectious bronchitis virus</td>
<td>NP_040831</td>
</tr>
<tr>
<td>GD03T0013</td>
<td>SARS coronavirus GD03T0013</td>
<td>AAT0463</td>
</tr>
<tr>
<td>PC4-127</td>
<td>SARS coronavirus PC4-127</td>
<td>AUA93318</td>
</tr>
<tr>
<td>PC4-137</td>
<td>SARS coronavirus PC4-137</td>
<td>AAV49720</td>
</tr>
<tr>
<td>Civet007</td>
<td>SARS coronavirus Civet007</td>
<td>AAT04646</td>
</tr>
<tr>
<td>A022</td>
<td>SARS coronavirus A022</td>
<td>AAV19631</td>
</tr>
<tr>
<td>GD01</td>
<td>SARS coronavirus GD01</td>
<td>AAP51227</td>
</tr>
<tr>
<td>GZ02</td>
<td>SARS coronavirus GZ02</td>
<td>AAS00013</td>
</tr>
<tr>
<td>CUHK-W1</td>
<td>SARS coronavirus CUHK-W1</td>
<td>AAPP1567</td>
</tr>
<tr>
<td>TOR2</td>
<td>SARS coronavirus TOR2</td>
<td>AAP1037</td>
</tr>
<tr>
<td>Urbani</td>
<td>SARS coronavirus Urbani</td>
<td>AAP13441</td>
</tr>
<tr>
<td>Frankfurt1</td>
<td>SARS coronavirus Frankfurt1</td>
<td>AAPP3697</td>
</tr>
<tr>
<td>Sino1-11</td>
<td>SARS coronavirus Sino1-11</td>
<td>AAR23250</td>
</tr>
</tbody>
</table>

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...
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

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


**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).