Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 203871, 8 pages
http://dx.doi.org/10.1155/2014/203871

Research Article

DV-Curve Representation of Protein Sequences and Its Application

Wei Deng[1,2] and Yihui Luan[1]

[1] *School of Mathematics, Shandong University, Jinan 250100, China*
[2] *School of Science, Shandong Jianzhu University, Jinan 250101, China*

Correspondence should be addressed to Yihui Luan; yhluan@sdu.edu.cn

Received 7 January 2014; Revised 10 March 2014; Accepted 3 April 2014; Published 8 May 2014

Academic Editor: Rui Jiang

Based on the detailed hydrophobic-hydrophilic(HP) model of amino acids, we propose dual-vector curve (DV-curve) representation of protein sequences, which uses two vectors to represent one alphabet of protein sequences. This graphical representation not only avoids degeneracy, but also has good visualization no matter how long these sequences are, and can reflect the length of protein sequence. Then we transform the 2D-graphical representation into a numerical characterization that can facilitate quantitative comparison of protein sequences. The utility of this approach is illustrated by two examples: one is similarity/dissimilarity comparison among different ND6 protein sequences based on their DV-curve figures the other is the phylogenetic analysis among coronaviruses based on their spike proteins.

1. Introduction

The graphical representation method has become very common to analyze the huge amount of gene data. Generally, with this method we can first observe visual qualitative inspection in order to recognize major differences among similar gene sequences and further draw some mathematical characterizations of sequences to analyze their similarity/dissimilarity and evolutionary homology.

Letter sequence representation (LSR) of DNA sequences represents each base by a letter of four different letters such as A, T, G, and C. DNA sequences can be represented in different dimension spaces. For example, G-curve and H-curve [1] were first proposed by Hamori and Ruskin before thirty years. Later, Gates [2] established a 2D graphical representation that was simpler than H curve. However, Gate's graphical representation has high degeneracy because of some circuits appearing in its curve. Several researchers in their recent studies have outlined different kinds of DNA sequences graphical representation based on 2D [3–11], 3D [12–15], 4D [16], 5D [17], and 6D [18]. Among these methods, we here stress DV-curve representation which was proposed by Zhang [10]. DV-curve uses two vectors to represent one alphabet of DNA sequences and avoids degeneracy and loss of information. Furthermore, DV-curve has good visualization no matter how long these sequences are and can reflect the length of the DNA sequence.

LSR of protein sequences represents each amino acid by a letter of twenty different letters such as A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, and V. Although protein sequences and DNA sequences belong to symbolic sequences, the methods for the graphical representation of protein sequences are relatively less popular, compared with DNA sequences. The key reason is that the extension of DNA graphical representation to protein sequences enormously increases the number of possible alternative assignments for these 20 amino acids. The amino acid sequence is the key to discover protein structure and function in the cell, so analysis of amino acid sequences is a very important part of postgenomic studies. The graphical representation study of protein sequences emerged very recently. The first visualization protein model was proposed by Randić et al. until 2004 [19]. Some researchers have studied on graphical representation of protein sequences from different perspectives [20–29].

In this paper, we introduce DV-curve graphical representation of protein sequences based on the detailed hydrophobic-hydrophilic (HP) model of amino acids. According to the important hydropathy, this approach is accompanied by a relatively small number of arbitrary choices associated with the graphical representation of proteins. Also, this representation has relatively good visualization effect to describe protein sequences in a perceivable way. As its application, we analyze the similarity/dissimilarity among some ND6 sequences and construct the phylogenetic tree of 35 coronavirus spike proteins.

2. DV-Curve Representation of Protein Sequences

2.1. Classification of Protein Sequences. The amino acid sequence is closely related to biological function. The closer the genetic relationship is, the smaller the difference in amino acid composition between them will be. Over the past thirty years, the characteristics of protein sequences have been studied by establishing different classified models [21–24, 26, 27]. A well-known model of protein sequences is the hydrophobic (H or nonpolar)-hydrophilic (P or polar), that is, the HP model may be too simple and lacks enough consideration on the heterogeneity and the complexity of the natural set of residues [30]. Based on Brown's work [31], 20 different kinds of amino acids are divided into four groups: nonpolar (np), negative polar (nep), uncharged polar (up), and positive polar (pp). This is called the detailed HP model, which can provide more information than the original HP model.

For a given protein sequence $S = S_1S_2\cdots S_n$ with length n, where S_i is the letter in the ith position among the protein sequence ($i = 1, 2, \ldots, n$), we define a primary protein sequence as a symbolic sequence which includes four letters according to the following rule:

$$b_i = \begin{cases} B_1, & \text{if } S_i \in \text{np}, \\ B_2, & \text{if } S_i \in \text{nep}, \\ B_3, & \text{if } S_i \in \text{up}, \\ B_4, & \text{if } S_i \in \text{pp}. \end{cases} \tag{1}$$

So b_i is the substitution for S_i, and then we obtain a sequence $G(s) = b_1b_2\cdots b_n$. Here b_i is a letter of the alphabet B_1, B_2, B_3, B_4. For example, for a given protein primary sequence $S = WTFESRNDPAK$, we can transform it into a new sequence according to the above rule, $G(S) = B_1B_3B_1B_2B_3B_4B_3B_2B_1B_1B_4$. Via comparison of the reduced sequence, it will be easier to understand the biological function of various kinds of amino acid residues.

2.2. Graphical Representation of Protein Sequences. In this section, we will construct DV-curve representation of protein sequence. Given any protein primary sequence with length n, we can transform it into a new sequence composed of a character set of B_1, B_2, B_3, B_4. As shown in Figure 1, these alphabets are assigned, respectively, by consecutive vectors as follows:

$$\begin{aligned} B_1 &\Longrightarrow (1,1),(1,1) \\ B_2 &\Longrightarrow (1,1),(1,-1) \\ B_3 &\Longrightarrow (1,-1),(1,1) \\ B_4 &\Longrightarrow (1,-1),(1,-1). \end{aligned} \tag{2}$$

We connect adjacent dots with lines and then obtain a dual-vector curve form. This process is shown in Figure 2.

Based on the construction of DV-curve, we obtain two mathematical models, respectively. One is "from protein sequence to DV-curve," and the other is "from DV-curve to protein sequence." Firstly, we give some common symbols and variables. (1) According to the classification rule, we describe a protein sequence as $G(S) = b_1b_2b_3\cdots b_n$, where $b_i \in \{B_1, B_2, B_3, B_4\}$ with length n. It means that the protein sequence S is connected by these alphabets. (2) (x_i, y_i) is the coordinate of the ith point of DV-curve, and $(x_0, y_0) = (0, 0)$ is the start point.

Model One. Given a primary protein sequence, we can draw its DV-curve:

$$\begin{aligned} x_{2i-1} &= 2i-1, \quad i = 1, 2, \ldots, n, \\ x_{2i} &= 2i, \quad i = 1, 2, \ldots, n, \\ y_{2i-1} &= \begin{cases} y_{2i-2} + 1, & \text{if } b_i = B_1 \text{ or } B_2, \\ y_{2i-2} - 1, & \text{if } b_i = B_3 \text{ or } B_4, \end{cases} \\ y_{2i} &= \begin{cases} y_{2i-1} + 1, & \text{if } b_i = B_1 \text{ or } B_3, \\ y_{2i-1} - 1, & \text{if } b_i = B_2 \text{ or } B_4. \end{cases} \end{aligned} \tag{3}$$

According to the above four formulas, the coordinate of each point (x_i, y_i) can be calculated. Then we connect all the points with beelines, and the DV-curve is obtained.

Model Two. Given a DV-curve, we can also obtain the coarse-grained description of the protein sequence based on the detailed HP-model:

$$G(S_i) = \begin{cases} B_1, & \text{if } y_{2i-1} - y_{2i-2} = 1, \quad y_{2i} - y_{2i-1} = 1, \\ B_2, & \text{if } y_{2i-1} - y_{2i-2} = 1, \quad y_{2i} - y_{2i-1} = -1, \\ B_3, & \text{if } y_{2i-1} - y_{2i-2} = -1, \quad y_{2i} - y_{2i-1} = 1, \\ B_4, & \text{if } y_{2i-1} - y_{2i-2} = -1, \quad y_{2i} - y_{2i-1} = -1. \end{cases} \tag{4}$$

Here $i = 1, 2, 3, \ldots, n$. If each point (x_i, y_i) of DV-curve is given in this model, we can get each B_i according to the above formulas. So the simplified protein sequence $G(S) = b_1b_2\cdots b_n$ can be recovered; here $b_i \in \{B_1, B_2, B_3, B_4\}$ with length n.

3. Numerical Characterization of Protein Sequences

In order to facilitate quantitative comparisons of sequences, we will give numerical characterization of graphical curve as

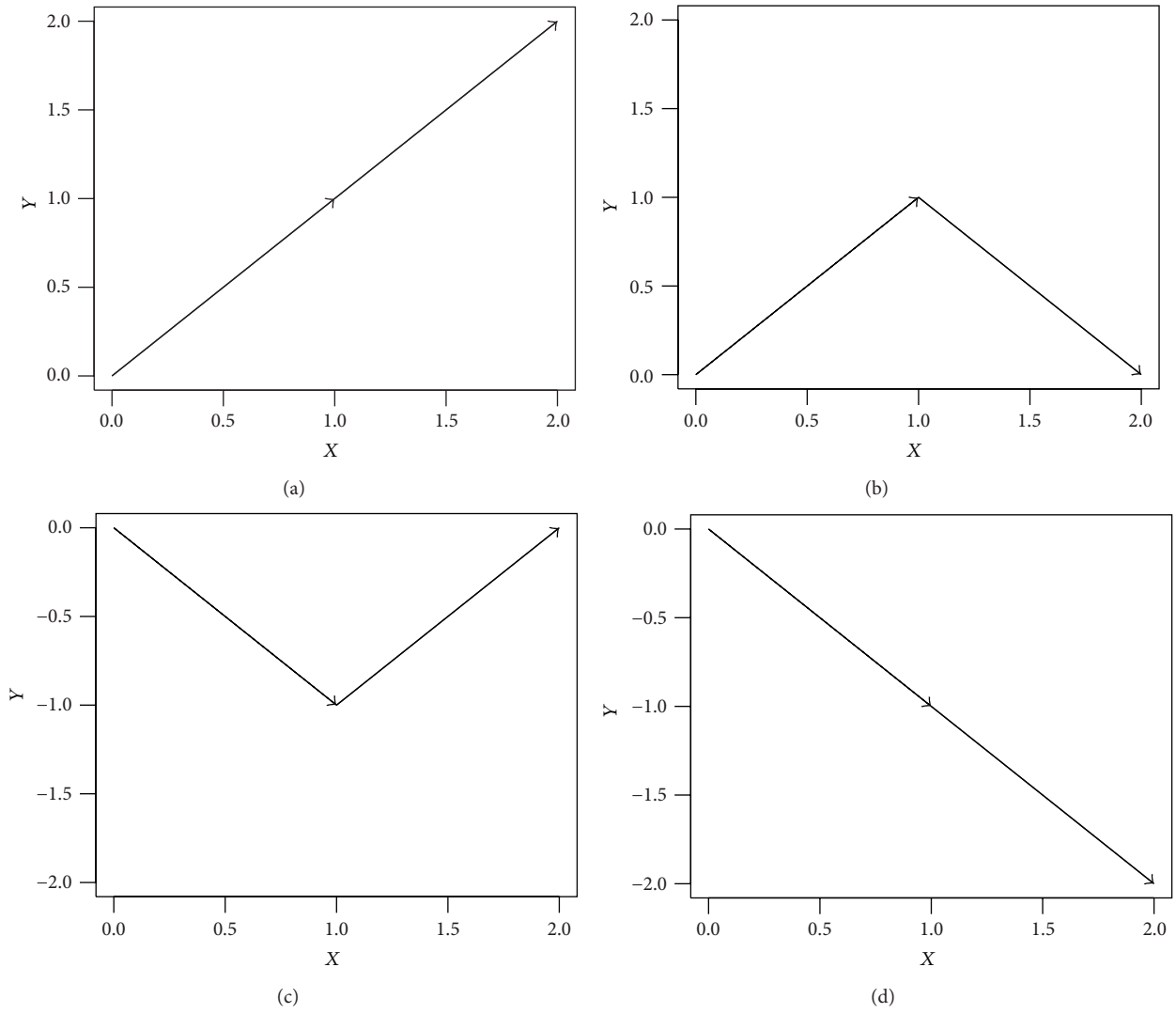

Figure 1: The representation of four alphabets of DV-curve: (a) B_1, (b) B_2, (c) B_3, and (d) B_4.

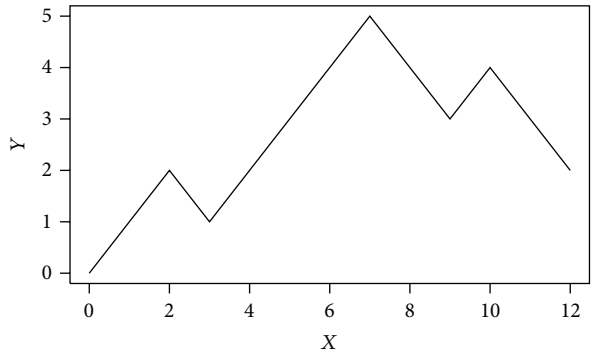

Figure 2: The DV-curve of sequence "WTFESR."

the descriptor. In general, we transform the graphical representation into a mathematical object like a matrix in order to draw some invariants. The frequently used matrices include E matrix, M matrix, L matrix, and L^k matrix proposed by Randić et al. [6, 8, 32–34]. Of course, there are some other matrix invariants such as the average matrix element, the average row sum, the Wiener number, and the ALE-index et al. These methods were used widely and proved to be useful. Here, we use the CM_{xy} as an alternative sequence invariant proposed by Liao et al. [35]:

$$(x_c, y_c) = \left(\frac{1}{2n+1} \sum_{i=0}^{2n} x_i,\ \frac{1}{2n+1} \sum_{i=0}^{2n} y_i \right),$$
$$CM_{xy} = \frac{1}{2n+1} \sum_{i=0}^{2n} (x_i - x_c)(y_i - y_c). \quad (5)$$

Obviously, this index is relatively simple for calculation so that this index can provide some convenience for long sequences.

If we adjust the order of B_1, B_2, B_3, B_4 corresponding to basic dual vectors, we can get another curve. So for a given sequence, we can get 4! = 24 different DV-curves totally. Therefore, a protein primary sequence can

Table 1: The information of 35 coronavirus spike proteins.

Number	Accession number	Abbreviation notation	Length (aa)	Group
1	P10033	FCoV1	1452	I
2	Q66928	FCoV2	1454	I
3	Q91AV1	PEDV3	1383	I
4	Q9DY22	TGEV4	1449	I
5	P18450	TGEV5	1449	I
6	P36300	CCoV6	1451	I
7	Q9J3E7	MHV7	1324	II
8	Q83331	MHV8	1361	II
9	P11224	MHV9	1324	II
10	O55253	MHV10	1360	II
11	Q9IKD1	RtCoV11	1360	II
12	P25190	BCoV12	1363	II
13	P15777	BCoV13	1363	II
14	Q9QAR5	BCoV14	1363	II
15	P36334	BCoV15	1363	II
16	P36334	HCoV16	1353	II
17	Q82666	IBV17	1166	III
18	P05135	IBV18	1163	III
19	P12722	IBV19	1154	III
20	Q64930	IBV20	1168	III
21	Q82624	IBV21	1159	III
22	P11223	IBV22	1162	III
23	Q98Y27	IBV23	1162	III
24	AAP41037	SCoV24	1255	IV
25	AAP300030	SCoV25	1255	IV
26	AAR91586	SCoV26	1255	IV
27	AAP51227	SCoV27	1255	IV
28	AAP33697	SCoV28	1255	IV
29	AAP13441	SCoV29	1255	IV
30	AAQ01597	SCoV30	1255	IV
31	AAU81608	SCoV31	1255	IV
32	AAS00003	SCoV32	1255	IV
33	AAR86788	SCoV33	1255	IV
34	AAR23250	SCoV34	1255	IV
35	AAT76147	SCoV35	1255	IV

be characterized by a 24-component vector as follows: $\vec{v} = [CM1_{xy}, CM2_{xy}, \ldots, CM24_{xy}]$. Based on the vectors, we can compare different protein sequences. Generally speaking, we can obtain the similarities of the two vectors by calculating Euclidean distance. If two sequences are similar, the distance between two corresponding points should be small. Given two species i and j, the corresponding vectors are $\vec{v}_i = [CMi1_{xy}, CMi2_{xy}, \ldots, CMi24_{xy}]$ and $\vec{v}_j = [CMj1_{xy}, CMj2_{xy}, \ldots, CMj24_{xy}]$, respectively; then we have $d(\vec{v}_i, \vec{v}_j) = \sqrt{\sum_{k=1}^{24}(CMik_{xy} - CMjk_{xy})^2}$.

4. Application

The comparison on biology sequences is one of the most important parts in bioinformatics when analyzing similarities of function and properties. In this section, we will give two main applications of this new graphical representation. One is similarity analysis based on visual graphics. Generally, similarity analysis can be divided into two types of methodologies to conduct the comparison: sequence alignment and sequence descriptors comparison. When recognizing figures, our brain is more helpful for similarity analysis in multiple sequences. So it is desirable to propose similarity analysis by inspecting the DV-curve of protein. The other is evolutionary homology analysis based on the numerical characterization of DV-curve, and we construct a 24-component vector to characterize any protein sequence. As further work, the phylogenetic tree of 35 coronavirus spike proteins is constructed.

4.1. Similarity Analysis Based on Visual Inspection of the Protein DV-Curve Graphs. Since Smith and Waterman developed a dynamic programming algorithm in 1981, many alignment algorithms identifying whether two biological sequences are similar to each other have been studied. These methods are proved to be efficient. However, multiple sequence alignment (MSA) of several hundred sequences has always produced a bottleneck.

In 1994, MSA was proved to be an NP-complete problem by Wang and Jiang [36]. Moreover, most experts think that it is impossible until now to build a deterministic polynomial algorithm to handle an NP-complete problem. It needs to exhaust almost billions or trillions of years. Except long computational time, there also exists possible bias of multiple sequence alignments for multiple occurrences of highly similar sequence [37].

However, our brain is much more powerful than computer when recognizing different figures. So it can help us to analyze the similarity in multiple sequences. If we can provide a simple, intuitional, clear, and nondegenerate 2D graphical representation of protein sequences, molecular biologists may easily find out which sequence is most similar or dissimilar to the given target sequence. And next they can use alignment algorithms for further confirmation.

According to our proposed definition of protein DV-curve, we can draw the curves of some ND6 (NADH dehydrogenase subunit 6) proteins in order to conveniently compare them. Protein sequences that are used to prove our approach were downloaded from GenBank: human (*YP*_003024037.1), gorilla (*NP*_008223), chimpanzee (*NP*_008197), wallaroo (*NP*_007405), harbor seal (H. seal) (*NP*_006939), gray seal (G. seal) (*NP*_007080), rat (*AP*_004903), and mouse (*NP*_904339), and the same data set was used in [26, 27].

In Figure 3, it is evident that protein graph of wallaroo is obviously different from the other species because it is the most remote species from the remaining mammals. Furthermore, we can see human and chimpanzee have similar curves, harbor seal and gray seal's curves are almost identical, and two curves of rat and mouse are very similar. All these results not only are consistent with the conclusions drawn by Smith-Waterman algorithm, but also agree well with the known fact of evolution and results drawn by other authors [26, 27, 38–40]. In particular, compared with the conclusion of [27], the DV-curve representation reflecting the similarities of sequences is more simple, intuitional, and visible.

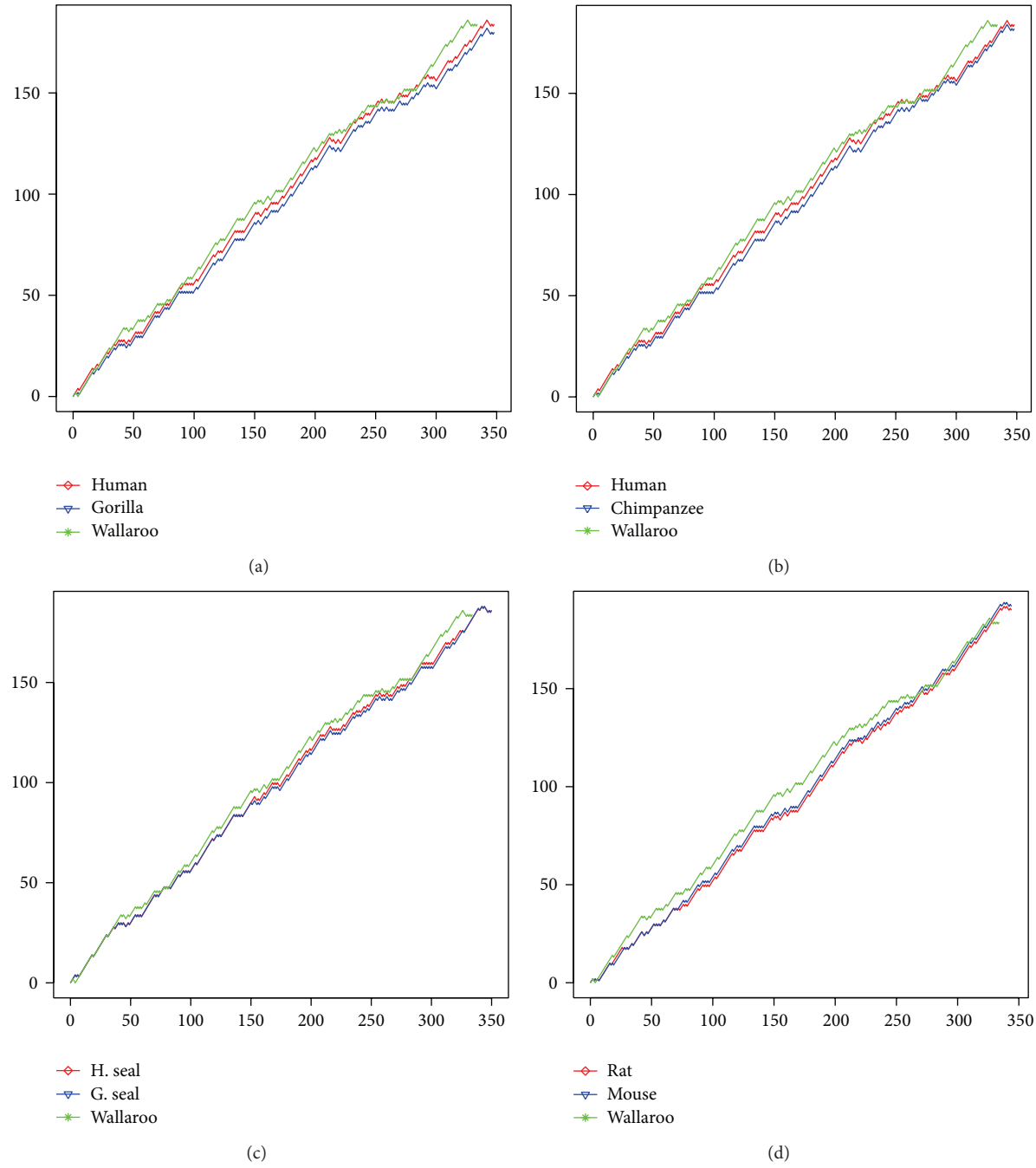

Figure 3: The DV-curve graphical representations of different ND6 proteins.

4.2. The Phylogenetic Analysis among the Spike Glycoprotein of Coronaviruses. Coronaviruses belong to order Nidovirales, family Coronaviridae, and genus *Coronavirus*. They are a diverse group of large, enveloped, single-stranded RNA viruses that cause respiratory and enteric diseases in humans and other animals. Generally, coronaviruses can be divided into three groups: the first group and the second group come from mammalian; the third group comes from poultry (chicken and turkey). A novel coronavirus has been identified as the cause of the outbreak of severe acute respiratory syndrome (SARS). Previous phylogenetic analysis based on sequence alignments shows that SARS-CoVs come from a new group distantly related to the above three groups of previously characterized coronaviruses [41, 42]. The spike (S) protein, which is common to all known coronaviruses, is crucial for viral attachment and entry into the host cell. To illustrate the use of DV-curve of protein sequences, we will construct the phylogenetic tree of 35 coronavirus spike proteins of Table 1.

As we have described above, a protein sequence can be associated with a 24-component vector. Given two species i and j, we can calculate the distance between them. Our

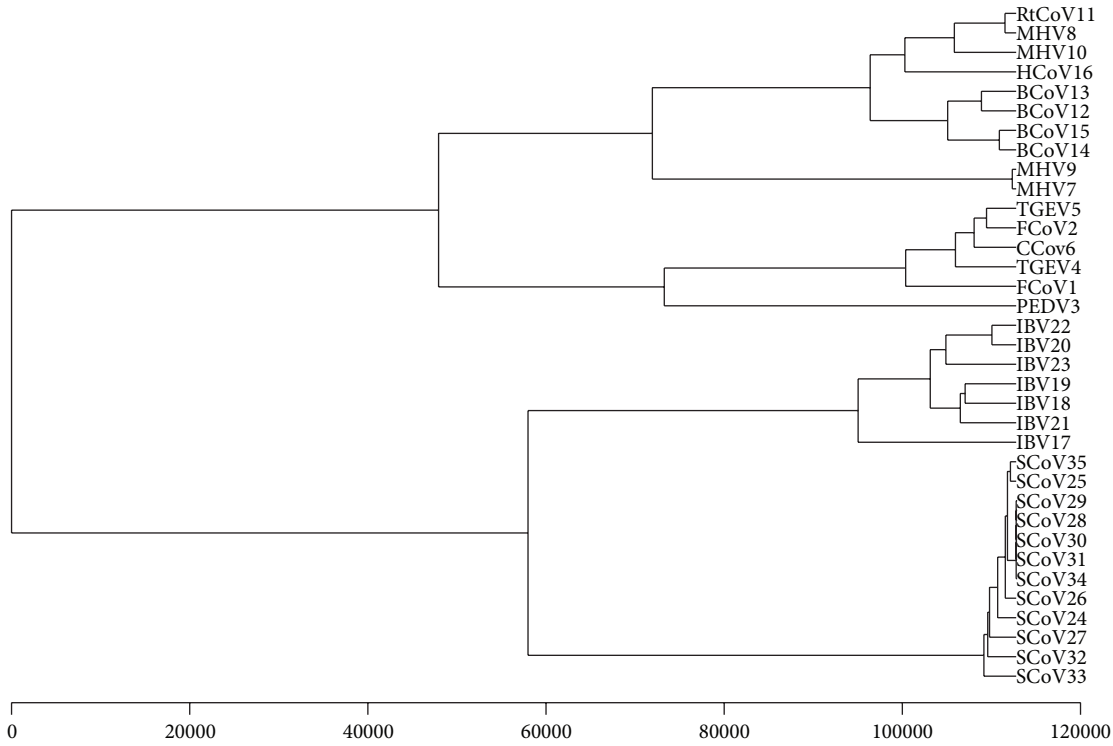

Figure 4: The phylogenetic tree based on the spike proteins.

datasets used in this paper were downloaded from GenBank (see Table 1 for details). Corresponding to 35 spike proteins, a 35×35 real symmetric matrix $D = (d_{ij})$ is obtained and used to reflect the evolutionary distance of them. Using the UPGMA program included in PHYLIP package 3.65, we can construct the phylogenetic tree of these 35 species [43, 44]. The branch lengths are not scaled according to the distances and only the topology of the tree is concerned.

Figure 4 shows coronaviruses can be overall divided into four groups. Furthermore, it is evident that SARS-CoVs appear to cluster together and form a separate branch, which can be distinguished easily from the other three groups of coronaviruses.

RtCoV11, MHV8, MHV10, HCoV16, BCoV13, BCoV12, BCoV15, BCoV14, MHV9, and MHV7, which belong to group 2, are situated at an independent branch, while TGEV5, FCoV2, CCoV6, TGEV4, FCoV1, and PEDV3, belonging to group 1, tend to cluster together. Meanwhile, the group 3 coronaviruses, including IBV22, IBV20, IBV23, IBV19, IBV18, IBV21, and IBV17, tend to cluster together in another branch. The resulting monophyletic clusters agree well with the established taxonomic groups [45, 46]. The conclusion is similar to that reported by other authors [23, 24]. Compared with result [24], it is noteworthy that a closer look at the subtree of the first branch shows coronavirus from three different species; that is, MHV, BCoV, and HCoV can be separated clearly, while they cluster together in a subtree by Li's method. Obviously, our conclusion is more consistent with the known evolution fact.

5. Conclusion

According to the detailed hydrophobic-hydrophilic (HP) model of amino acids, we can reduce a protein primary sequence containing 20 amino acids into a four-letter sequence, which can be treated as a coarse-grained description of the protein primary sequence. Here we cannot avoid losing some information in the reduced sequences, but we can focus our main attention on the part of our interest.

Some alignment-free methods to analyze DNA sequences have been proposed. However, there are few alignment-free methods to analyze protein sequences. Our method realizes the generalization from DNA graphical representations to those of proteins acceptable and can be seen a valid supplement to graphical representation of protein sequences. Meanwhile we first propose to combine DV-curve and the detailed HP model together to describe protein sequences.

Compared with classical Smith-Waterman algorithm, the similarity/dissimilarity analysis results are consistent with DV-curve. In addition, the advantage of our method is that it can visualize the local and global features among different proteins no matter how long these sequences are and avoid degeneracy at the same time. The new approach is applied in two aspects: one is similarity intuitive analysis of ND6 protein sequences of several species and the other is phylogenetic analysis among 35 coronaviruses based on their spike proteins. Results have shown that our proposed method is more intuitional, simple, effectual, and feasible.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank to all the anonymous reviewers for their valuable suggestions and support. This research is supported by the National Science Foundation of China Grants 11371227 and 10921101.

References

[1] E. Hamori and J. Ruskin, "H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences," *Journal of Biological Chemistry*, vol. 258, no. 2, pp. 1318–1327, 1983.

[2] M. A. Gates, "A simple way to look at DNA," *Journal of Theoretical Biology*, vol. 119, no. 3, pp. 319–328, 1986.

[3] A. Nandy, "Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences," *Computer Applications in the Biosciences*, vol. 12, no. 1, pp. 55–62, 1996.

[4] X. F. Guo, M. Randic, and S. C. Basak, "A novel 2-D graphical representation of DNA sequences of low degeneracy," *Chemical Physics Letters*, vol. 350, no. 1-2, pp. 106–112, 2001.

[5] A. Nandy and P. Nandy, "On the uniqueness of quantitative DNA difference descriptions in 2D graphical representation models," *Chemical Physics Letters*, vol. 368, no. 1-2, pp. 102–107, 2003.

[6] M. Randic, M. Vracko, N. Lers, and D. Plavsic, "Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation," *Chemical Physics Letters*, vol. 371, pp. 202–207, 2003.

[7] Y. H. Yao and T.-M. Wang, "A class of new 2-D graphical represent ation of DNA sequences and their application," *Chemical Physics Letters*, vol. 398, no. 4–6, pp. 318–323, 2004.

[8] M. Randic, "Graphical representations of DNA as 2-D map," *Chemical Physics Letters*, vol. 386, pp. 468–471, 2004.

[9] G. H. Huang, B. Liao, Y. F. Li, and Z. B. Liu, "H-L curve: a novel 2D graphical representation for DNA sequences," *Chemical Physics Letters*, vol. 462, no. 1–3, pp. 129–132, 2008.

[10] Z.-J. Zhang, "DV-Curve: a novel intuitive tool for visualizing and analyzing DNA sequences," *Bioinformatics*, vol. 25, no. 9, pp. 1112–1117, 2009.

[11] W. Deng and Y. H. Luan, "Analysis of similarity/dissimilarity of DNA sequences based on chaos game representation," *Abstract and Applied Analysis*, vol. 2013, Article ID 926519, 6 pages, 2013.

[12] B. Liao and K. Ding, "A 3D graphical representation of DNA sequences and its application," *Theoretical Computer Science*, vol. 358, no. 1, pp. 56–64, 2006.

[13] Z. Cao, B. Liao, and R. Li, "A group of 3D graphical representation of DNA sequences based on dual nucleotides," *International Journal of Quantum Chemistry*, vol. 108, no. 9, pp. 1485–1490, 2008.

[14] Y. J. Huang and T. M. Wang, "New graphical representation of a DNA sequence based on the ordered dinucleotides and its application to sequence analysis," *International Journal of Quantum Chemistry*, vol. 112, no. 6, pp. 1746–1757, 2012.

[15] B. Liao, Y. S. Zhang, K. Q. Ding, and T.-M. Wang, "Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation," *Journal of Molecular Structure: THEOCHEM*, vol. 717, no. 1–3, pp. 199–203, 2005.

[16] R. Chi and K. Ding, "Novel 4D numerical representation of DNA sequences," *Chemical Physics Letters*, vol. 407, no. 1–3, pp. 63–67, 2005.

[17] B. Liao, R. Li, W. J. Zhu, and X. Xiang, "On the similarity of DNA primary sequences based on 5-D representation," *Journal of Mathematical Chemistry*, vol. 42, no. 1, pp. 47–57, 2007.

[18] B. Liao and T.-M. Wang, "Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 5, pp. 1666–1670, 2004.

[19] M. Randić, J. Zupan, and A. T. Balaban, "Unique graphical representation of protein sequences based on nucleotide triplet codons," *Chemical Physics Letters*, vol. 397, no. 1-3, pp. 247–252, 2004.

[20] F. L. Bai and T. M. Wang, "A 2-D graphical representation of protein sequences based on nucleotide triplet codons," *Chemical Physics Letters*, vol. 413, no. 4–6, pp. 458–462, 2005.

[21] N. Liu and T. M. Wang, "Protein-based phylogenetic analysis by using hydropathy profile of amino acids," *FEBS Letters*, vol. 580, no. 22, pp. 5321–5327, 2006.

[22] M. Randić, "2-D Graphical representation of proteins based on physico-chemical properties of amino acids," *Chemical Physics Letters*, vol. 440, no. 4–6, pp. 291–295, 2007.

[23] C. Li, L. L. Xing, and X. Wang, "2-D graphical representation of protein sequences and its application to coronavirus phylogeny," *Journal of Biochemistry and Molecular Biology*, vol. 41, no. 3, pp. 217–222, 2008.

[24] D. D. Li, J. Wang, and C. Li, "New 3-D graphical representation of protein sequences and its application," *China Journal of Bioinformatics*, vol. 7, no. 1, pp. 60–63, 2009.

[25] J. Wen and Y. Zhang, "A 2D graphical representation of protein sequence and its numerical characterization," *Chemical Physics Letters*, vol. 476, no. 4–6, pp. 281–286, 2009.

[26] Y. H. Yao, Q. Li, N. Li, X. Y. Nan, P. A. He, and Y. Z. Zhang, "Similarity/dissimilarity studies of protein sequences based on a new 2d graphical representation," *Journal of Computational Chemistry*, vol. 31, no. 5, pp. 1045–1052, 2010.

[27] X.-L. Xie, L.-F. Zheng, Y. Yu et al., "New technique: protein sequence analysis based on hydropathy profile of amino acids," *Journal of Zhejiang University: Science B*, vol. 13, no. 2, pp. 152–158, 2012.

[28] M. I. Abo El Maaty, M. M. Abo-Elkhier, and M. A. Abd Elwahaab, "3D graphical representation of protein sequences and their statistical characterization," *Physica A: Statistical Mechanics and Its Applications*, vol. 389, no. 21, pp. 4668–4676, 2010.

[29] M. M. Abo-EIkhier, "Similarity/dissimilarity analysis of protein sequences using the spatial median as a descriptor," *Journal of Biophysical Chemistry*, vol. 3, no. 2, pp. 142–148, 2012.

[30] J. Wang and W. Wang, "Modeling study on the validity of a possibly simplified representation of proteins," *Physical Review E*, vol. 61, no. 6, pp. 6981–6986, 2000.

[31] T. A. Brown, *Genetics*, Chapman & Hall, London, UK, 3rd edition, 1998.

[32] M. Randić, M. Vračko, A. Nandy, and S. C. Basak, "On 3-D graphical representation of DNA primary sequences and their numerical characterization," *Journal of Chemical Information and Computer Sciences*, vol. 40, no. 5, pp. 1235–1244, 2000.

[33] M. Randic, M. Vracko, L. Nelia, and P. Dejan, "Novel 2-D graphical representation of DNA sequences and their numerical characterization," *Chemical Physics Letters*, vol. 368, no. 1-2, pp. 1–6, 2003.

[34] M. Randic, M. Vracko, J. Zupan, and M. Novic, "Compact 2-D graphical representation of DNA," *Chemical Physics Letters*, vol. 373, pp. 558–562, 2003.

[35] B. Liao, M. Tan, and K. Ding, "Application of 2-D graphical representation of DNA sequence," *Chemical Physics Letters*, vol. 414, pp. 296–300, 2005.

[36] L. Wang and T. Jiang, "On the complexity of multiple sequence alignment," *Journal of Computational Biology*, vol. 1, no. 4, pp. 337–348, 1994.

[37] T. D. Pham and J. Zuegg, "A probabilistic measure for alignment-free sequence comparison," *Bioinformatics*, vol. 20, no. 18, pp. 3455–3461, 2004.

[38] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang, "An information-based sequence distance and its application to whole mitochondrial genome phylogeny," *Bioinformatics*, vol. 17, no. 2, pp. 149–154, 2001.

[39] H. H. Otu and K. Sayood, "A new sequence distance measure for phylogenetic tree construction," *Bioinformatics*, vol. 19, no. 16, pp. 2122–2130, 2003.

[40] V. Makarenkov and F.-J. Lapointe, "A weighted least-squares approach for inferring phylogenies from incomplete distance matrices," *Bioinformatics*, vol. 20, no. 13, pp. 2113–2121, 2004.

[41] T. G. Ksiazek, S. R. Zaki, C. Urbani et al., "A novel coronavirus associated with severe acute respiratory syndrome," *The New England Journal of Medicine*, vol. 348, pp. 1953–1966, 2003.

[42] M. A. Marra, S. J. Jones, C. R. Astell et al., "The genome sequence of the sars-associated coronavirus," *Science*, vol. 300, p. 1399, 2003.

[43] P. H. A, R. R. Sneath, and Sokal, *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, 1973.

[44] PHILIP, http://evolution.gs.washington.edu/phylip.html.

[45] P. A. Rota, M. S. Oberste, S. S. Monroe et al., "Characterization of a novel coronavirus associated with severe acute respiratory syndrome," *Science*, vol. 300, no. 5624, pp. 1394–1399, 2003.

[46] S. K. P. Lau, P. C. Y. Woo, K. S. M. Li et al., "Severe acute respiratorysyndrome coronavirus-like virus in Chinese horseshoe bats," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 39, pp. 14040–14045, 2005.