A Sequence-Segmented Method Applied to the Similarity Analysis of Proteins

Fen Kong[1], Xu-ying Nan[2], Ping-an He[1], Qi Dai[2], Yu-hua Yao[*2]
[1]College of Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China
[2]College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China
zaozuang1989@126.com, nanxuying@163.com, pinganhe@zstu.edu.cn, daiailiu2004@yahoo.com.cn, yaoyuhua2288@163.com

Abstract—***A 2-D graphical representation of protein sequences based on two classifications of amino acids is outlined. The method of dividing a long sequence into k segments (SSM) is introduced, so protein graph is divided into k segments, geometrical center of the points for all protein curve segment is given as descriptors of protein sequences. It is not only useful for comparative study of proteins, but also for encoding innate information about the structure of proteins. Finally, a simple example is taken to highlight the behavior of the new descriptor on protein sequences taken from the 12 baculoviruse proteins.***

Keywords-Similarity; Sequence-Segmented Method; Graphical representation; Descriptors.

I. INTRODUCTION

Bio-molecular sequence comparison is the origin of bioinformatics. Today, powerful sequence comparison methods, together with comprehensive biological databases, have changed the practice of molecular biology and genomics. Previously, almost all such comparisons are based on sequence alignment: these methods use dynamic programming, a score function is used to represent insertion, deletion, and substitution of nucleotides or amino acids in the compared DNAs or proteins, finally a regression technique that finds an optimal alignment by assigning scores to different possible alignments and picking the alignment with the highest score. Recently, biological sequence analysis quickly incorporated additional concepts and algorithms, such as stochastic modeling of sequences using hidden Markov models and other Bayesian theory methods for hypothesis testing and parameter estimation [1].

Among all existing alignment-free methods for comparing biological macromolecules, graphical representation techniques provide a simple way to view, sort, and compare sequences or structures. H-curve, graphical representation of DNA sequences was introduced by Hamori in 1983 [2]. Graphical representations of bio-sequences were expanded from DNA [3-5], RNA secondary stucture [6, 7] to proteins [8, 9] and as it grew from qualitative and pictorial representations to quantitative estimation of sequence similarities/ dissimilarities. These graphical representations both 2-D and 3-D can be associated with a matrix, such as E, M/M, L/L, ${}^{k}L/{}^{k}L$, thus the matrix invariants arrive at various numerical descriptors rather than the visual description of sequence. The comparison of sequences changed into the comparison of descriptors. Above matrix methods by forming ratios of graph theoretic and Euclidean distances between nodes of the graphical plots, first formulated for DNA sequences in Randic et al. Those methods are used in the study of global homology and conserved patterns, the analysis of similarity and dissimilarity, the study of fractal and long range correlations. This technique has been widely used method of choice for the researchers in this field who have defined different types of matrices to construct various invariants for describe the bio-sequences. However, the difficulties associated with computing various parameters for very large matrices that are natural for large sequences have restricted the numerical characterizations to leading eigenvalues and the like [10].

Another approach using geometrical descriptor was proposed by Raychaudhury and Nandy [11], and it has been found to be useful for several calculations based on the 2D graphical representation [12], and extended recently to an abstract 20D modelling for protein sequences [13], where individual sequences are indexed by numerical descriptors. The approach is convenient, fast and efficient, but it couldn't used to similarity/dissimilarity measure for bio-sequences with length less than 1000.

In this paper, we outlined a dynamic 2-D graphical representation based two physico-chemical properties of amino acids, and introduced a novel strategy for sequence comparison based on the method of dividing a long sequence into k segments (SSM). We will make a comparison for helicase protein sequences of 12 baculoviruses, including 3 group I NPVs: AcMNPV (*Autographa californica* MNPV), BmNPV (*Bombyx mori* NPV), RoMNPV (*Rachiplusia ou* MNPV); 5 group II NPVs: HearNPV (*Helicoverpa armigera* NPV), HzSNPV (*Helicoverpa zea* SNPV), MacoNPVA (*Mamestra configurata* NPVA), MacoNPVB (*Mamestra configurata* NPVB), SeMNPV (*Spodoptera exigua* MNPV); 3 GVs: AdorGV (*Adoxophyles orona* GV), CpGV (*Cydia pomonella* GV), CrleGV (*Cryptophlebia leucotreta* GV); 1 hymenopteran baculovirus: NeseNPV (*Neodiprion sertifer* NPV). The family baculoviridae is divided into two genera, Nucleopolyhedrovirus (NPV) and Granulovirus (GV). Lepidopteran NPVs show a further division into group I and group II NPVs. Group I NPVs appear to be much more conserved than those of group II [14]. Length and group information of these protein sequences are showed in Table 1. The similarities are computed by calculating the Euclidean distance among the end point of the normalized descriptor vectors. Using our approach, one can find that the

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

computational complexity is only $O(N)$, and greatly reduces the computational complexity.

II. OUTLINE THE DYNAMIC 2-D GRAPHICAL REPRESENTATION OF PROTEINS

Here we consider two physic-chemical properties which have important relations with structure of proteins: chirality and hydrophilicity of 20 amino acids. The two sets of physicochemical characteristic of 20 amino acids are listed in Table 2. In the following chapters, we will construct the 2-D graphical representations of protein sequences. The two properties of amino acids, cirality and hydrophobicity which can be selected as the basis for construct two dimensional Cartesian coordinates. In Figure 1, we show the 2-D map of amino acids resulting from ordering the amino acids along the x-axis with respect to cirality and along the y-axis with respect to hydrophobicity.

TABLE I. LENGTH AND GROUP INFORMATION OF HELICASE PROTEIN SEQUENCES OF 12 BACULOVIRUSE

baculoviruse	group information	length
AcMNPV	Group I NPVs	1221
BmNPV		1222
RoMNPV		1221
HearNPV	Group II NPVs	1253
HzSNPV		1253
MacoNPVA		1212
MacoNPVB		1209
SeMNPV		1222
AdorGV	GVs	1138
CpGV		1131
CrleGV		1128
NeseNPV	Hymenopteran NPV	1143

First, enantiomeric molecules display a special property called chirality (or optical activity)—the ability to rotate the plane of polarization of plane-polarized light. Clockwise rotation of incident light is referred to as dextrorotatory behavior, and counterclockwise rotation is called levorotatory behavior. The magnitude and direction of the optical rotation depend on the nature of the amino acid side chain. Based on the chirality of amino acids for H2O, 20 amino acids are simplified into 3 types: dextrorotatory amino acids D={R, D, Q, E, A, I, K, V}; levorotatory amino acids L={N, C, H, L, M, F, P, S, T, W}; and irrotational (irrational) amino acids I={G, Y} (because tyrosine is not soluble in water). Accordingly, we denote that:

$$x_i = \begin{cases} +1 & \text{if } S_i \in D = \{R, D, Q, E, A, I, K, V\}; \\ 0 & \text{if } S_i \in I = \{G, Y\}; \\ -1 & \text{if } S_i \in L = \{N, C, H, L, M, F, P, S, T, W\}. \end{cases}$$

Second, the hydrophobicity of amino acids is an important property. In a protein, hydrophobic amino acids are likely to be found in the interior, whereas hydrophilic amino acids are likely to be in contact with the aqueous environment. Based on their hydrophobicity, 20 amino acids are simplified into 3 types: hydrophobic amino acids H={C, M, F, I, L, V, W, Y}; hydrophilic amino acids P={N, Q, D, E, R, K, H}; and neutral amino acids N={A, G, T, P, S}. Accordingly, we denote that:

$$y_i = \begin{cases} +1 & \text{if } S_i \in H = \{C, M, F, I, L, V, W, Y\}; \\ 0 & \text{if } S_i \in N = \{A, G, T, P, S\}; \\ -1 & \text{if } S_i \in P = \{N, Q, D, E, R, K, H\}. \end{cases}$$

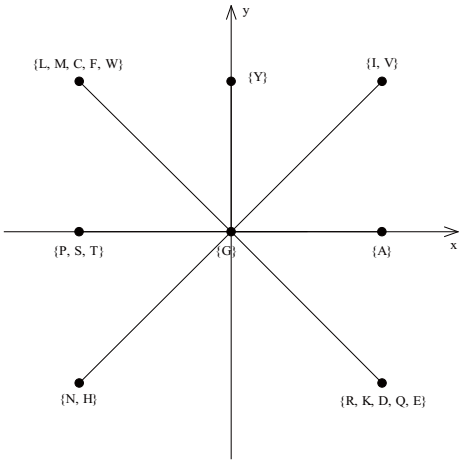

Figure 1. The 2-D map of 20 amino acids: twenty special vectors represent the twenty amino acids, respectively.

Thus, given a protein sequence $S = s_1 s_2 \cdots s_N$ with N amino acids, inspect it by stepping one amino acid at a time. For the step i $(i = 1,2,\cdots,N)$, a 2-D space point $P_i(x_i, y_i)$ can be constructed as follows:

$$(x_i, y_i) = (x_{i-1}, y_{i-1}) + \begin{cases} (0,0) & \text{if } S_i \in \{G\}; \\ (-1,0) & \text{if } S_i \in \{P, S, T\}; \\ (+1,0) & \text{if } S_i \in \{A\}; \\ (-1,-1) & \text{if } S_i \in \{N, H\}; \\ (+1,-1) & \text{if } S_i \in \{R, K, D, Q, E\}; \\ (0,+1) & \text{if } S_i \in \{Y\}; \\ (-1,+1) & \text{if } S_i \in \{L, M, C, F, W\}; \\ (+1,+1) & \text{if } S_i \in \{I, V\}. \end{cases}$$

where $(x_0, y_0) = (0,0)$. When i runs from 1 to N, we have points $P_1, P_2, \cdots, P_N$. Connecting adjacent points, we obtain a 2-D zigzag curve.

During the construction of the graph, we preset the value of properties corresponding to the positive and negative direction of the axis of coordinates. Actually, if we exchange the distribution of value +1 and -1 in one property, they are symmetry of one of the coordinate plane. So under the generalized symmetry the graph we obtain is unique. Obviously, amino acid Glycine (G) is an immobile dot in

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

graphical representation of protein sequence, but it has same effect with another 19 amino acids in similarity analysis.

We will illustrate the current approach on two shorter segments of a protein of yeast Saccharomyces cerevisiae. In Figure 2 we illustrate for two proteins zigzag curves, obtained by connecting adjacent amino acids using their vectors sequentially as posed in Figure 1. The corresponding proteins are:

Protein I: WTFESRNDPAKDPVILWLNGGPGCSSLTGL

Protein II: WFFESRNDPANDPIILWLNGGPGCSSFTGL

Observe Figure 2, two proteins zigzag curves of Protein I and Protein II are similar on the whole, and have several same local sequence's segments.

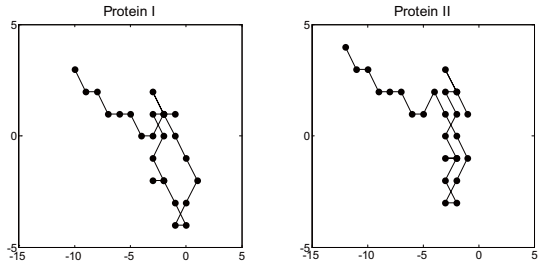

Figure 2. The dynamic 2-D graph of Protein I and Protein II.

Afterwards, the graphical representations of the 12 baculoviruse proteins for visualization are showed in Figure 3. Viewing the curves, we can find that the curves of 3 group I NPVs (AcMNPV, BmNPV, RoMNPV) are similar, the graphs of (HearNPV, HzSNPV), (MacoNPVA, MacoNPVB, SeMNPV) in 5 group II NPVs are similar, respectively. And 3 GVs (AdorGV, CpGV, CrleGV) are also similar. In addition, we find protein graph of NeseNPV is obviously different from other species. Their similarities/dissimilarities are consistent with classification of these baculoviruse proteins.

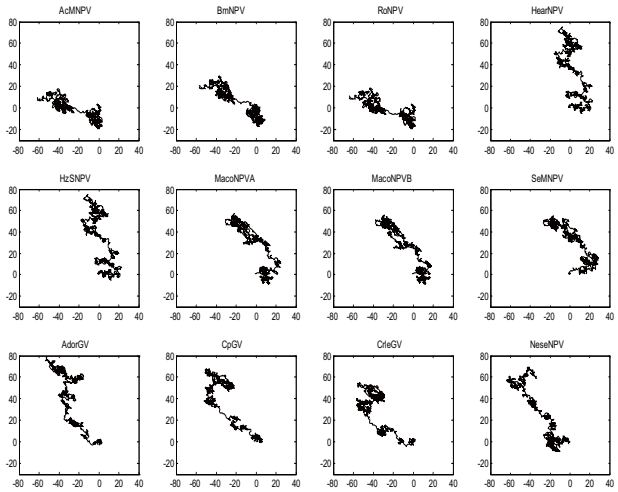

Figure 3. The graphical representations of 12 baculoviruse proteins.

III. NUMERICAL CHARACTERIZATION

A. Geometrical center

Once we can use some of matrix invariants as descriptors of the sequence. But, the computational complexity of these matrix invariants techniques is at least $O(N^2)$, which results in the main difficulty in computation. In this section, we bypass the difficulty and introduce two ways to numerically characterize protein sequence. Their computational complexities are reduced to $O(N)$, so it is easy to implement.

In the new model, the protein sequences are represented by a set of material points in 2-D space. In order to find some of the invariants sensitive to the form of the characteristic curve, we will transform the characteristic curve into another mathematical object. In the Cartesian coordinate axis systems, Nandy [11] denote

$$\begin{cases} \mu_x = \dfrac{1}{N}\sum_{i=1}^{N} x_i \\ \mu_y = \dfrac{1}{N}\sum_{i=1}^{N} y_i \end{cases}$$

as the geometrical center (a weighted mean of the coordinate values of the representative points) of the points corresponding protein curve and regard the geometrical center as the descriptors for the dynamic 2-D graph, where N represents the total length of the protein sequence, x_i and y_i are the coordinates of the i-th amino acid in the Cartesian coordinate system with the point (0, 0) as the origin of all the sequences. In Table 3, we illustrate the geometrical center of the dynamic 2-D graphs representing of 12 baculoviruse proteins.

Based on the results have got above, we construct 2-component vectors of the 2-D graphs corresponding to 12 baculoviruse proteins. We give the similarity/dissimilarity matrices for the 12 baculoviruse protein sequences based on the Euclidean distances between the 2-component vectors of the geometrical center in table 4. The results of the similarity are mainly consistent to the known fact of evolution. Most of the similarity values are consistent with classification of these baculoviruse proteins. That is to say, the geometrical centers may be more effective to numerically characterize protein sequences. Whereas, we found that the 4 baculoviruse proteins of NeseNPV, AdorGV, CpGV and CrleGV are more similar with each other. Unique 1 hymenopteran NPV, NeseNPV isn't separated from 3 GVs. This result is not consistent with the known conclusion of evolution. It is may cased by the loss of information in the process of graphical representation model, and this may be due to too long to the baculoviruse protein sequences.

B. New strategy

For overcome the difficulty that the geometrical center of protein graph is unfit for long sequence, we outlined a strategy:

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

the method of dividing a long protein sequence into k segments (SSM), length of each segment is

$$\overbrace{ceil(l/k),\cdots,ceil(l/k)}^{\mathrm{mod}(l,k)},\overbrace{floor(l/k),\cdots,floor(l/k)}^{k-\mathrm{mod}(l,k)},$$

respectively. In which, $\mathrm{mod}(l,k)$, divides l by k and returns a remainder that is a whole number, $floor(X)$ rounds the elements of X to the nearest integers towards minus infinity, $ceil(X)$ rounds the elements of X to the nearest integers towards infinity. For example, length of AcMNPV protein is 1221, take k=5, its curve is divided into 5 segments, length of each segment is 245, 244, 244, 244, 244, respectively.

Geometrical centers of k segments are $(\mu_{\mathrm{x}}^1,\mu_{\mathrm{y}}^1),(\mu_{\mathrm{x}}^2,\mu_{\mathrm{y}}^2),\cdots,(\mu_x^k,\mu_y^k)$, respectively. We propose to take a combined $2k$-dimension vector,

$$\vec{v}(S)=\left(\mu_{\mathrm{x}}^1,\mu_{\mathrm{y}}^1,\mu_{\mathrm{x}}^2,\mu_{\mathrm{y}}^2,\cdots\mu_x^k,\mu_y^k\right)$$

as the descriptors for the 2D-dynamic graph. In this paper, we take *k*=5, the 5 pairs geometrical centers of the dynamic 2-D graphs representing of 12 baculoviruse proteins are showed in Table 5.

IV. SIMILARITIES AND DISSIMILARITIES

Give two arbitrary sequences S^1 and S^2. In the graphical approaches, the respective 2*k*-dimensions vectors are composed for the geometrical centers corresponding to *k* segments of characteristic curves of S^1 and S^2. Such similarity/diversity comparisons of sequence S^1 and S^2 are based on Euclidean distance between the end points of two normalized vectors. The Euclidian distance $D(S^1,S^2)$ between the two vectors is

$$D(S^1,S^2)=\left\|\vec{v}(S_1)-\vec{v}(S_2)\right\|_2$$

The analysis of similarity/dissimilarity represented by the vectors is based on the assumption that two proteins are similar if their corresponding vectors point to a similar direction and have similar magnitudes. That is to say, the smaller the Euclidean distance is, the more similar the two proteins are. Based on the Euclidean distances between the 10-component vectors of the geometrical center, the similarity/dissimilarity matrices for the 12 baculoviruse protein sequences is represented in Table 6.

Observing Table 6, the smaller entries are associated with the pairs in group (AcMNPV, BmNPV, RoMNPV), (HearNPV, HzSNPV), (MacoNPVA, MacoNPVB, SeMNPV) and (AdorGV, CpGV, CrleGV). On the other hand, the larger entries in the similarity/dissimilarity matrix appear in the rows belonging to NeseNPV. These results are consistent with the known conclusion of evolution, and we think that it is not accidental.

V. CONCLUSION

It is well-known that the alignments of protein sequences are computer intensive that is direct comparison for alphabet sequences. Structure considered in alignments of sequences is only string's structures. In this paper,

TABLE II. THE TWO SETS OF PHYSICOCHEMICAL CHARACTERISTIC PARAMETERS OF 20 AMINO ACIDS

Amino acid	Symbol	Specific Rotation [a] [α]D(H2O)	hydrophobicity	x	y
Glycine	G	-	N	0	0
Alanine	A	+1.8	N	1	0
Proline	P	-86.2	N	-1	0
Valine	V	+5.6	H	1	1
Leucine	L	-11.0	H	-1	1
Isoleucine	I	+12.4	H	1	1
Methionine	M	-10.0	H	-1	1
Phenylalanine	F	-34.5	H	-1	1
Tyrosine	Y	-	H	0	1
Tryptophan	W	-33.7	H	-1	1
Serine	S	-7.5	N	-1	0
Threonine	T	-28.5	N	-1	0
Cysteine	C	-16.5	H	-1	1
Asparagine	N	-5.3	P	-1	-1
Glutamine	Q	+6.3	P	1	-1
Lysine	K	+13.5	P	1	-1
Histidine	H	-38.5	P	-1	-1
Arginine	R	+12.5	P	1	-1
Aspartate	D	+5.0	P	1	-1
Glutamate	E	+12.0	P	1	-1

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

TABLE III. *X*-COORDINATES OF THE CENTERS OF MASS (μ_x), *Y*-COORDINATES OF THE CENTERS OF MASS (μ_y) OF THE 2D-DYNAMIC GRAPHS REPRESENTING HELICASE PROTEIN SEQUENCES OF 12 BACULOVIRUSES.

baculoviruse	μ_x	μ_y
AcMNPV	-22.8239	-1.8812
BmNPV	-18.3502	5.3961
RoMNPV	-24.5741	-1.8452
HearNPV	2.4381	36.1189
HzSNPV	0.7007	36.1165
MacoNPVA	-2.4538	27.3449
MacoNPVB	-7.5203	25.5385
SeMNPV	2.7831	30.5475
AdorGV	-29.5431	43.3032
CpGV	-30.9752	39.9080
CrleGV	-36.4078	30.7943
NeseNPV	-29.6080	25.4357

TABLE IV. SIMILARITY/DISSMILARITY TABLE BASED ON GEOMETRICAL CENTER OF 2D GRAPH FOR THE HELICASE PROTEIN SEQUENCES OF 12 BACULOVIRUSES

	BmNPV	RoMNPV	HearNPV	HzSNPV	MacoNPVA	MacoNPVB	SeMNPV	AdorGV	CpGV	CrleGV	NeseNPV
AcMNPV	8.5424	1.7506	45.6310	44.6905	35.6245	31.4013	41.3200	45.6813	42.5769	35.3866	28.1468
BmNPV		9.5484	37.0952	36.1481	27.1007	22.8693	32.8514	39.5250	36.7487	31.1632	22.9853
RoMNPV			46.5933	45.6060	36.6247	32.2599	42.3994	45.4210	42.2411	34.7185	27.7415
HearNPV				1.7374	10.0456	14.5298	5.5821	32.7782	33.6276	39.2092	33.7800
HzSNPV					9.3216	13.3970	5.9457	31.0859	31.9021	37.4882	32.1357
MacoNPVA						5.3789	6.1386	31.4403	31.1658	34.1288	27.2213
MacoNPVB							11.4565	28.2947	27.5067	29.3618	22.0880
SeMNPV								34.7519	35.0321	39.1917	32.7921
AdorGV									3.6848	14.2687	17.8676
CpGV										10.6100	14.5368
CrleGV											8.6575

TABLE V. THE GEOMETRICAL CENTER ($\mu_x^i, \mu_y^i, i = 1, \cdots 5$) OF THE 2-D GRAPH SEGMENTS OF HELICASE PROTEIN SEQUENCES OF 12 BACULOVIRUSES

	μ_x^1	μ_y^1	μ_x^2	μ_y^2	μ_x^3	μ_y^3	μ_x^4	μ_y^4	μ_x^5	μ_y^5
AcMNPV	-3.3592	-6.9020	-4.4180	-11.9180	-24.7705	-1.8607	-38.3115	5.0410	-43.3402	6.2541
BmNPV	-0.5918	-6.5347	0.2367	-8.1592	-20.6762	6.7582	-33.2049	16.6721	-37.6639	18.3484
RoMNPV	-3.9959	-6.2653	-5.5615	-11.9795	-26.7705	-3.8607	-41.2869	5.6434	-45.3402	7.2541
HearNPV	10.3705	2.7012	9.0717	17.6733	-6.3785	41.7570	-2.0800	58.6680	1.1840	59.9800
HzSNPV	10.3705	2.7012	7.9841	17.6614	-8.3785	41.7570	-4.6880	58.6680	-1.8160	59.9800
MacoNPVA	8.9547	-0.9918	13.8683	11.3128	-2.3967	33.0289	-12.8182	45.1322	-19.9917	48.4256
MacoNPVB	8.3058	-1.0289	9.3223	9.7438	-8.3719	28.1694	-20.2521	43.4793	-26.6846	47.4191
SeMNPV	15.5633	8.5388	19.3551	17.1878	0.8934	35.9836	-5.0164	43.5369	-17.0000	47.6352
AdorGV	-14.3728	9.1447	-30.7895	28.7763	-33.9123	50.2895	-41.2775	68.4053	-27.4053	60.0837
CrleGV	-6.9692	8.0264	-36.2168	29.3363	-45.4513	50.9779	-37.2522	58.7212	-29.0929	52.6195
CpGV	-16.8628	4.8274	-43.0133	22.2655	-44.8673	38.6416	-40.9956	46.4800	-36.3200	41.8756
NeseNPV	-12.2140	1.7118	-18.6376	-0.4803	-24.5197	23.4629	-45.9605	44.4211	-46.8553	58.2895

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

TABLE VI. SIMILARITY/DISSMILARITY TABLE BASED ON GEOMETRICAL CENTER OF SEGMENED GRAPH FOR THE HELICASE PROTEIN SEQUENCES OF 12 BACULOVIRUSES

	BmNPV	RoMNPV	HearNPV	HzSNPV	MacoNPVA	MacoNPVB	SeMNPV	AdorGV	CpGV	CrleGV	NeseNPV
AcMNPV	21.7827	4.9334	112.4125	109.9900	85.7863	75.9930	95.3746	112.7336	106.7895	89.8049	73.8668
BmNPV		24.0568	91.4055	88.9015	64.4490	54.5384	74.9438	97.0636	92.6776	79.3496	59.3682
RoMNPV			114.7620	112.2210	88.1928	77.9223	98.0378	112.5947	106.6377	89.1382	72.9279
HearNPV				4.5809	32.4281	42.0127	31.2068	75.1364	79.5975	91.4795	81.9224
HzSNPV					30.2607	38.9415	30.3624	71.3715	75.7437	87.7173	77.9864
MacoNPVA						13.6628	17.0783	76.0251	78.9065	83.6447	64.4200
MacoNPVB							27.8057	70.3911	72.2653	74.6191	52.7692
SeMNPV								84.6046	86.9325	93.7632	77.8359
AdorGV									19.6978	36.9343	53.3865
CpGV										26.2365	55.5755
CrleGV											46.8677

(1) We present a 2-D representation of proteins based on two significant physico-chemical proprieties. The other chemical or physical properties of amino acids will also be useful to study and solve some bioinformatics problems. The advantage of our approach is that it allows visual inspection of data, helping in recognizing major similarities among different proteins.

(2) Under the generalized symmetry, the uniqueness and the simplicity of the outlined 2-D graphical representation and accompanying numerical characterization of proteins offer, in our view, an attempt in the comparative study of proteins.

(3) For the long protein sequence, the coordinates are easily computed, many schemes can be used to numerically characterize protein sequences, and the examination of similarities/dissimilarities among the helicase protein sequences of 12 baculoviruses illustrates the utility of the approach. Our method has the advantage of greatly reducing the computational complexity.

(4) Our approach gives numerical characterization of proteins by graphic representation and used to similarity analysis of 12 helicase proteins. And similarly, our approach also can be applied to analysis and compare complete gene sequences from corresponding graphical representation.

Although the segmented graphical representation can speed up similarity analysis, the applicability of this method is limited. Because the method is essentially a global method, it does not help in finding large chunks of insertions/ deletions. In future work, we will identifying the different results of representation by different methods or specified *k*, phylogenetic tree of these 12 proteins can improve the presentation, the advantage of the presenting method will be much more clear.

ACKNOWLEDGMENT

We appreciate the financial support of this work that was provided by Zhejiang Provincial Natural Science Foundation of China (No. LY12F02043, No. Y1110752). This work was also partially supported by the National Natural Science Foundation of China (61170316, 61170110).

REFERENCES

[1] E. Hamori, J. Ruskin, “H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences,” J. Biol. Chem., 1983, vol. 258, pp.1318–1327.

[2] M. A. Gates. “A simple way to look at DNA,” J. Theor. Biol., 1986, vol. 119, pp. 319–328.

[3] P. M. Leong, S. Mogenthaler, walk and gap plots of DNA sequences. Compu.t Appl. Biosci., 1995, vol. 12, pp. 503–511.

[4] A. Nandy, A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. Curr. Sci., 1994, vol. 66, pp. 309–314.

[5] B. Liao, T. M. Wang, A 3D Graphical representation of RNA secondary structure, J. Biomol. Str. Dyn., 2004, vol. 21, pp. 827-832.

[6] 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. 2005, vol. 26, pp. 1339–1346.

[7] M. Randić, J. Zupan, A. T. Balaban, “Unique graphical representation of protein sequences based on nucleotide triplet codons,” Chem. Phys. Lett., 2004, vol. 397, pp. 247–252.

[8] M. Randić, “2-D Graphical representation of proteins based on virtual genetic code,” SAR QSAR Environ. Res., 2004, vol. 15, pp. 147–157.

[9] S. Vinga, J. Almeida, “Alignment-free sequence comparison—a review,” Bioinformatics, 2003, vol. 19, pp. 513–523.

[10] A. Ghosh, A. Nandy, P. Nandy, “Computational analysis and determination of a highly conserved surface exposed segment in H5N1 avian flu and H1N1 swine flu neuraminidase,” BMC Struct. Biol., 2010, 10:6.

[11] A. Nandy, P. Nandy, “On the uniqueness of quantitative DNA difference descriptors in 2D graphical representation models,” Chem. Phys. Lett., 2003, 368, (1-2), pp. 102–107.

[12] Y. H. Yao, X. Y. Nan, T. M. Wang, “Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation,” Chem. Phys. Lett., 2005, 411, (1-3), pp. 248–255.

[13] J. F. Yu, X. Sun, J. H. Wang, “A Novel 2D Graphical Representation of Protein Sequence Based on Individual Amino Acid,” International Int. J. Quantum Chem., 2011, 111, (12), pp. 2835–2843.

[14] E. A. Herniou, J. A. Olszewski, D. R. O'Reilly, J. S. Cory, “Ancient Coevolution of Baculoviruses and Their Insect Hosts,” J. Virol., 2004, 78, (7), pp. 3244–3251.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE