MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **70** (2013) 431-450

ISSN 0340 - 6253

A Sequence-Segmented Method Applied to the Similarity Analysis of Long Protein Sequence

Yu-hua Yao [1,*], Fen Kong [2], Qi Dai [1], Ping-an He [2]

[1] *College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, P.R. China*

[2] *College of Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, P.R. China*

(Received June 27, 2012)

Abstract: A 2-D graphical representation of protein sequences based on two classifications of amino acids is outlined. We transform the characteristic graphs into numerical characterization and used for similarity analysis of proteins. The method of dividing a long protein sequence into segments (SSM) is introduced, so protein graph is divided into *k* segments, geometrical center of the points for all protein curve segments is given as descriptors of proteins. It is not only useful for comparative study of proteins, but also for encoding amino acids in ways that the visualization of protein sequences facilitates the decoding of its information content. In addition, a simple example applied to the helicase proteins of 12 baculoviruses is taken to highlight the behavior of the new strategy.

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

* Corresponding author e-mail: yaoyuhua2288@163.com

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 [3-32], RNA secondary stucture [33-40] to proteins including protein sequence [41-49], protein secondary structure [50] and proteome [51-54], and as it grew from qualitative and pictorial representations to quantitative estimation of sequence similarities/dissimilarities [31,32].

These graphical representations both 2-D and 3-D can be associated with a matrix, such as E, L/L, ${}^{k}L\backslash{}^{k}L$, thus the matrix invariants arrive at various numerical descriptors rather than the visual description of sequence. The comparison of sequences is changed into the comparison of descriptors. Above matrix methods by forming the quotient between the Euclidean distances between vertices (atoms) *i* and *j* and the distance between the same two vertices when measured along the connecting bonds, first formulated for DNA sequences in Randić et al [30]. Those methods are used in the study of global homology and conserved patterns, the analysis of similarity/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. The method based on this descriptor has been improved and many other descriptors were proposed [55-58]. However, the difficulties associated with computing various parameters for very large matrices that are natural for long sequences have restricted the numerical characterizations to leading eigenvalues and the like [59].

Another method using geometrical descriptor of the curve was proposed by Raychaudhury and Nandy [60], and it has been found to be useful for several calculations based on the graphical representation of DNA sequences [19,20], and extended recently to mathematical descriptors for protein sequences [41,61,62]. The approach is convenient, fast and efficient, but it couldn't be used to similarity/dissimilarity measure for the long sequences with length large than 1000 [20,61].

In this paper, we outlined a new 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 protein sequence into k segments (SSM). We will make a comparison for helicase protein sequences of 12 baculoviruses, including 3 group I alphabaculovirus: AcMNPV, BmNPV, RoMNPV; 5 group II Alphabaculovirus: HearNPV, HzSNPV, MacoNPVA, MacoNPVB, SeMNPV; 3 betabaculovirus: AdorGV, CpGV, CrleGV; 1 gammabaculovirus (hymenopteran baculovirus): NeseNPV. The family abaculovirus 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 [63,64]. 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 computational complexity is only $O(N)$, and greatly reduces the computational complexity.

Table 1. Length and group information of helicase protein sequences of 12 baculoviruses

Clade (Group)	Virus name	Abbreviation	Accession No.	Length
Group I Alphabaculovirus	*Autographa californica* MNPV	AcMNPV	AAA66725	1221
	Bombyx mori NPV	BmNPV	AAC63764	1221
	Rachiplusia ou MNPV	RoMNPV	AAN28013	1221
Group II Alphabaculovirus	*Helicoverpa armigera* NPV	HearNPV	AEN04007	1253
	Helicoverpa zea SNPV	HzSNPV	AAL56093	1253

	Mamestra configurata NPVA	MacoNPVA	AAM09201	1212
	Mamestra configurata NPVB	MacoNPVB	AAM95079	1209
	Spodoptera exigua MNPV	SeMNPV	AAB96630	1222
Betabaculovirus	*Adoxophyles orona* GV	AdorGV	AAP85713	1138
	Cydia pomonella GV	CpGV	AAK70750	1131
	Cryptophlebia leucotreta GV	CrleGV	AAQ21676	1128
Gammabaculovirus	*Neodiprion sertifer* NPV	NeseNPV	AAQ96438	1143

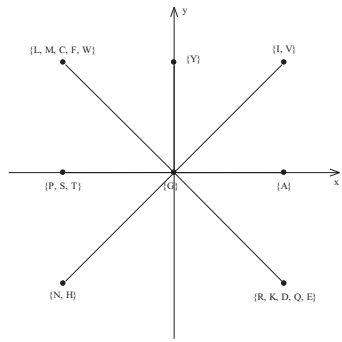

Figure 1. The 2-D map of 20 amino acids.

Outline the 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. 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 2-D 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.

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 H_2O, 20 amino acids are simplified into 3 types: dextrorotatory amino acids *D*={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: $x_i : D \to +1, I \to 0, L \to -1.$ 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, twenty 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: $y_i : H \to +1, N \to 0, P \to -1.$

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. Obviously, amino acid Glycine (G) is an immobile dot in 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. 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.

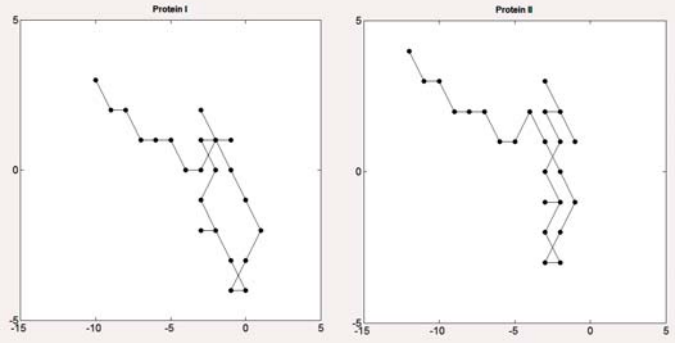

Figure 2. The 2-D characteristic graphs of Protein I and Protein II.

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 [63-67].

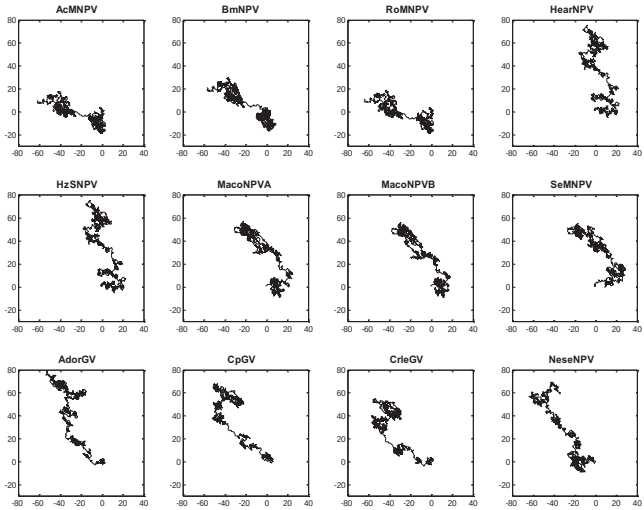

Figure 3. The graphical representations of helicase proteins of 12 baculoviruse.

Numerical characterization

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.

Geometrical center

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 [68] denote

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

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 2, we illustrate the geometrical center of the 2-D characteristic graphs representing of 12 helicase proteins.

Table 2. Geometrical center of the 2-D graphs

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

Based on the geometrical center, we construct 2-component vectors of the 2-D graphs corresponding to 12 baculoviruse proteins. In table 3, we give the similarity/dissimilarity matrices for the 12 helicase protein sequences based on the Euclidean distances between the 2-component vectors. 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: (1) among the entries of Table 3, the entries of (SeMNPV, HearNPV) and (SeMNPV, HzSNPV) are

smaller than that of (SeMNPV, MacoNPVA) and (SeMNPV, MacoNPVB), that is to say, SeMNPV is more similar to HearNPV and HzSNPV, in fact, SeMNPV, MacoNPVA and MacoNPVB are more similar with each other; (2) the 4 baculoviruse proteins of NeseNPV, AdorGV, CpGV and CrleGV are more similar with each other. Unique 1 hymenopteran NPV, NeseNPV hasn't separated from 3 GVs. These results are not consistent with the known conclusion of evolution. It is may cased by the loss of information in the process of graphical representation model.

Table 3. Similarity/Dissmilarity table based on geometrical center of 2D graph

Baculoviruse	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

New strategy

For overcome the difficulty that the geometrical center of protein graph is unfit for long sequence, we outlined a strategy: 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 $(\mu_x^1,\mu_y^1),(\mu_x^2,\mu_y^2),\cdots,(\mu_x^k,\mu_y^k)$, respectively. We propose to take a combined $2k$-dimension vector,

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

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 helicase proteins are showed in Table 4.

Table 4. The segments geometrical centers of the 2-D graph, k=5

Baculoviruse	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

Similarities/Dissimilarities of 12 helicase proteins

Give two arbitrary sequences S^1 and S^2. In the graphical approaches, the respective $2k$-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 helicase protein sequences for 12 baculoviruse are represented in Table 5 (k=5).

Table 5. Similarity/Dissmilarity table based on k segmented geometrical centers of 2D graph, k=5

Baculoviruse	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

Table 6. Similarity/Dissmilarity table based on k segmented geometrical centers of 2D graph, k=3

Baculoviruse	BmNPV	RoMNPV	HearNPV	HzSNPV	MacoNPVA	MacoNPVB	SeMNPV	AdorGV	CpGV	CrleGV	NeseNPV
AcMNPV	16.7186	3.7223	86.3257	84.4308	66.1801	58.5696	73.5534	85.3338	79.9301	66.5628	56.5309
BmNPV		18.4597	70.1172	68.1466	49.7406	42.0135	57.7669	73.0054	68.7136	58.0714	45.1719
RoMNPV			88.1296	86.1435	68.0469	60.0678	75.5867	85.2471	79.7228	65.9303	55.7900
HearNPV				3.5845	24.3508	32.0870	23.1632	57.0915	60.4774	69.5753	62.7335
HzSNPV					22.4827	29.5647	22.2779	54.0904	57.4102	66.5886	59.5944
MacoNPVA						10.6289	13.0437	56.3835	58.2133	62.1038	48.9851
MacoNPVB							21.1818	52.1005	52.9129	54.9406	39.8157
SeMNPV								62.7098	64.1886	70.0060	59.4765
AdorGV									13.8088	27.2700	38.6390
CpGV										20.0122	39.6585
CrleGV											32.3322

Table 7. Similarity/Dissmilarity table based on k segmented geometrical centers of 2D graph, k=20

Baculoviruse	BmNPV	RoMNPV	HearNPV	HzSNPV	MacoNPVA	MacoNPVB	SeMNPV	AdorGV	CpGV	CrleGV	NeseNPV
AcMNPV	43.9409	10.2894	228.3840	223.5672	174.6589	155.2459	194.7247	231.0309	220.4897	188.4448	153.1007
BmNPV		48.6406	186.7620	181.8037	132.4811	113.0994	154.4676	200.1712	192.8873	168.2553	125.0281
RoMNPV			233.1197	228.0723	179.6273	159.2276	200.1963	230.8529	220.3140	187.2966	151.4430
HearNPV				9.3984	68.1947	86.8928	67.7509	155.4051	166.2664	189.8059	168.2604
HzSNPV					63.8173	80.6324	65.7822	147.8996	158.5361	182.2712	160.3496
MacoNPVA						28.1450	38.8565	155.7227	163.3186	172.9910	132.0293
MacoNPVB							58.7627	144.1944	149.5198	154.7875	108.5394
SeMNPV								173.8508	178.9700	192.7148	159.7863
AdorGV									46.0946	77.6464	112.3755
CpGV										54.7710	119.1446
CrleGV											103.1305

For the parameter k, segmented number, we take k=3 and k=20, based on the Euclidean distances between the 2k-component vectors of the geometrical centers, the similarity/dissimilarity matrices for the 12 baculoviruse protein sequences is showed in Table 6 and Table 7, respectively.

Observing Tables 5-7, 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 [63-67].

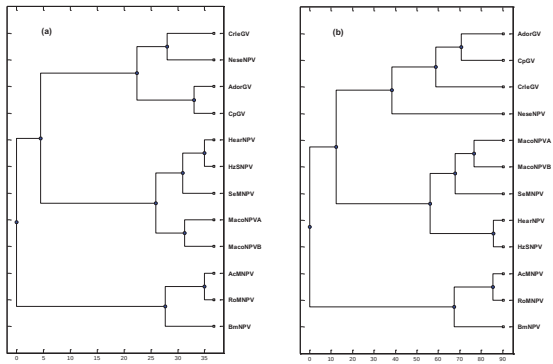

Figure 4. Two phylogenetic tree of 12 baculovirus, (a) geometrical center, (b) SSM: k=5.

In Figure 4, using the UPGMA method in the MATLAB Bioinformatics tool box, we gave the two phylogenetic trees of helicase protein sequences for 12 baculoviruse based on two distance matrices of Table 3 and Table 5. In Figure 4 (a), NeseNPV, AdorGV, CpGV and CrleGV are dislplayed a branch, in particular, NeseNPV showed a high similarity to CrleGV; SeMNPV, HearNPV and HzSNPV are dislplayed a branch. In Figure 4 (b), NeseNPV and 3 GVs are split; MacoNPVA, MacoNPVB and SeMNPV are displayed a branch. Clearly, the results of Figure 4 (b) are consistent with the known conclusion of evolution.

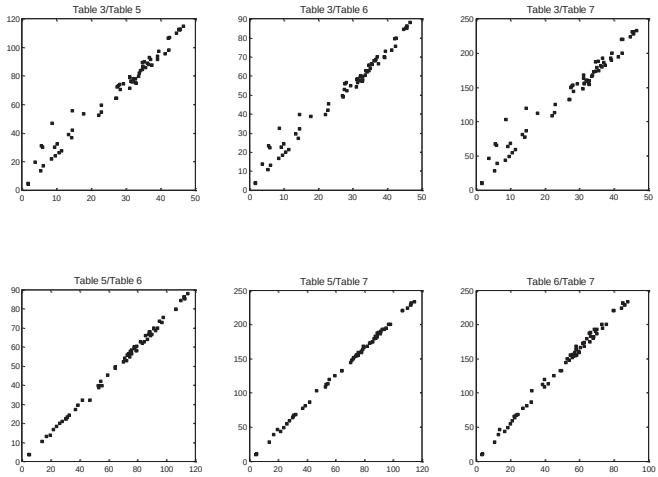

Figure 5. The correlations of the Table 3 and Tables 5-7.

In addition, we can consider the entries of similarity/dissimilarity tables and correlate entries of one table with the corresponding entries of another table. Figure 5 presented the correlations for Tables 3/Table5, Table 3/Table 6, Tables 3/Table 7, Table 5/Table 6, Table 5/Table 7 and Table 6/Table 7, in which the x coordinate is shown as numerator and the y coordinates as the denominator. The correlations among the similarity/dissimilarity Tables mainly have the following three cases: (1) the worst

correlation occurs in Table 3 and other Tables, because the points in Table 3/Table5, Table 3/Table 6 and Tables 3/Table 7 are very dispersive. It suggests that there is inconsistent information obtained by the geometrical center of graph and SSM, this result support the conclusion that sequence segment and no segment are different; (2) Table 5/Table 6 and Table 6/Table 7 shows the better proportionality, which means that Tables 5-7 carry similar information (and support each other) but they are still slightly different, hence encode slightly different structural features; (3) the best correlation occurs in the Table 5/Table 7, which shows the influence of the parameter k i.e. sequence-segmented number between 5-20 is slight.

Conclusion and Discussion

It is well-known that the alignments of protein sequences are computer intensive that is direct comparison for alphabet sequences. The multiple alignment strategy does not work for all types of data, e.g. whole genome phylogeny, and the evolutionary models may not always be correct. Structure considered in alignments of sequences is only string's structures. In this paper,

(1) We present a 2-D graphical representation of protein sequences based on two significant physicochemical proprieties. The advantage of our approach is that it allows visual inspection of data, helping in recognizing major similarities among different proteins. 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.

(2) For the long protein sequence, the coordinates are easily computed, many schemes can be used to numerically characterize protein sequences, and the examination of similarity/dissimilarity illustrates the utility of the approach. Our method doesn't require alignment, one can find that the computational complexity is only $O(N)$, and greatly reducing the computational complexity of protein sequence comparison.

(3) Our approach gives numerical characterization of proteins by graphic representation and used to analyze the similarity of 12 helicase proteins. Also, both

computational scientists and molecular biologists can use it to analysis protein sequences efficiently. We have divided the 20 amino acids in terms of two physic-chemical properties, chirality and hydrophobicity, to plot their two-dimensional graphs. One could also have used other properties of the amino acids such as their basic or acidic nature.

(4) In the similarity analysis of 12 helicase proteins, the influence of the parameter *k* i.e. sequence-segmented number is take 5-20 is slight. It suggests that there is inconsistent information obtained by the geometrical center of graph and SSM, hence encode slightly different structural features. This result support the conclusion that sequence segment and no segment are different.

(5) Although the segmented graphical representation can speed up similarity analysis, the applicability of this method is limited. Would the results drawn from the SSM prescription differ if we were to use such other properties? What would be the way to choose which properties to use for any particular task? How to select a value for the parameter *k* for different bioinformatics problems? This would be important if we are to use this method to understand new protein sequences where other evolutionary information may not be available. In future work, we will identify the different results of representation by different methods or specified *k*.

Acknowledgment: We appreciate the financial support of this work that was provided by the National Natural Science Foundation of China (No. 61272312, No. 61170316 and No. 61170110). This work was also supported by Zhejiang Provincial Natural Science Foundation of China (No. LY12F02043, No. Y1110752).

References

[1] S. Vinga, J. Almeida, Alignment-free sequence comparison-a review, *Bioinformatics* **19** (2003) 513–523.

[2] E. Hamori, J. Ruskin, H-curves, a novel method of representation of nucleotide series especially suited for long DNA-sequences, *J. Biol. Chem.* **258** (1983) 1318–1327.

[3] D. Bielinska-Wąż, Four-component spectral representation of DNA sequences, *J. Math. Chem.* **47** (2010) 41–51.

[4] Q. Dai, X. Q. Liu, T. M. Wang, A novel 2D graphical representation of DNA sequences and its application, *J. Mol. Graph.* **25** (2006) 340–344.

[5] R. Jayalakshmi, R. Natarajan, M. Vivekanandan, Extension of molecular similarity analysis approach to classification of DNA sequences using DNA descriptors, *SAR QSAR Environ. Res.* **22** (2011) 21–34.

[6] C. Li, J. Wang, Numerical characterization and similarity analysis of DNA sequences based on 2-D graphical representation of the characteristic sequences, *Comb. Chem. High Throughput Screen.* **6** (2003) 795–799.

[7] B. Liao, K. Q. Ding, A 3D graphical representation of DNA sequences and its application, *Theor. Comput. Sci.* **358** (2006) 56–64.

[8] B. Liao, C. Zeng, F. Q. Li, Y. Tang, Analysis of similarity/dissimilarity of DNA sequences based on dual nucleotides, *MATCH Commun. Math. Comput. Chem.* **59** (2008) 647–652.

[9] B. Liao, W. Zhu, Y. Liu, 3D graphical representation of DNA sequence without degeneracy and its applications in constructing phylogenic tree, *MATCH Commun. Math. Comput. Chem.* **56** (2006) 209–216.

[10] Z. B. 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.

[11] A. Nandy, Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences, *Comput. Appl. Biosci.* **12** (1996) 55–62.

[12] A. Nandy, S. C. Basak, New approaches to drug-DNA interactions based on graphical representation and numerical characterization of DNA sequences, *Curr. Comput.-Aided Drug Des.* **6** (2010) 283–289.

[13] Z. H. Qi, T. R. Fan, PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization, *Chem. Phys. Lett.* **442** (2007) 434–440.

[14] M. Randić, Another look at the chaos-game representation of DNA, *Chem. Phys. Lett.* **456** (2008) 84–88.

[15] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, *Chem. Phys. Lett.* **371** (2003) 202–207.

[16] M. Randić, J. Zupan, Highly compact 2D graphical representation of DNA sequences, *SAR QSAR Environ. Res.* **15** (2004) 191–205.

[17] M. Randić, J. Zupan, D. Vikić-Topić, D. Plavšić, A novel unexpected use of a graphical representation of DNA: Graphical alignment of DNA sequences, *Chem. Phys. Lett.* **431** (2006) 375–379.

[18] Y. H. Yao, Q. Dai, X. Y. Nan, P. A. He, Z. M. Nie, S. P. Zhou, Y. Z. Zhang, Analysis of similarity/dissimilarity of DNA sequences based on a class of 2D graphical representation, *J. Comput. Chem.* **29** (2008) 1632–1639.

[19] 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.* **411** (2005) 248–255.

[20] Y. H. Yao, X. Y. Nan, T. M. Wang, A new 2D graphical representation – Classification curve and the analysis of similarity/dissimilarity of DNA sequences, *J. Mol. Struct. (Theochem)* **764** (2006) 101–108.

[21] Y. H. Yao, T. M. Wang, A class of new 2-D graphical representation of DNA sequences and their application, *Chem. Phys. Lett.* **398** (2004) 318–323.

[22] C. L. Yu, M. Deng, S. S. T. Yau, DNA sequence comparison by a novel probabilistic method, *Inf. Sci.* **181** (2011) 1484–1492.

[23] J. F. Yu, J. H. Wang, X. Sun, Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation, *MATCH Commun. Math. Comput. Chem.* **63** (2010) 493–512.

[24] Y. S. Zhang, A simple method to construct the similarity matrices of DNA sequences, *MATCH Commun. Math. Comput. Chem.* **60** (2008) 313–324.

[25] Y. S. Zhang, W. Chen, Invariants of DNA sequences based on 2DD-curves, *J. Theor. Biol.* **242** (2006) 382–388.

[26] Y. S. Zhang, W. Chen, New invariant of DNA sequences, *MATCH Commun. Math. Comput. Chem.* **58** (2007) 197–208.

[27] M. A. Gates, A simple way to look at DNA, *J. Theor. Biol.* **119** (1986) 319–328.

[28] A. Nandy, Graphical representation of long DNA-sequences, *Curr. Sci.* **66** (1994) 821–821.

[29] P. M. Leong, S. Morgenthaler, Random walk and gap plots of DNA sequence, *Comput. Appl. Biosci.* **11** (1995) 503–507.

[30] M. Randić, M. Vračko, A. Nandy, S. C. Basak, On 3-D graphical representation of DNA primary sequences and their numerical characterization, *J. Chem. Inf. Comput. Sci.* **40** (2000) 1235–1244.

[31] M. Novič, M. Randić, Representation of proteins as walks in 20-D space, *SAR QSAR Environ. Res.* **19** (2008) 317–337.

[32] A. Nandy, A. Ghosh, P. Nandy, Numerical characterization of protein sequences and application to voltage-gated sodium channel alpha subunit phylogeny, *In Silico. Biol.* **9** (2009) 77–87.

[33] Y. S. Zhang, On 2D graphical representation of RNA secondary structure, *MATCH Commun. Math. Comput. Chem.* **57** (2007) 697–710.

[34] Y. S. Zhang, On 3D graphical representation of RNA secondary structure, *MATCH Commun. Math. Comput. Chem.* **57** (2007) 157–168.

[35] Y. Zhang, J. Q. Qiu, L. Q. Su, Comparing RNA secondary structures based on 2D graphical representation, *Chem. Phys. Lett.* **458** (2008) 180–185.

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

[37] J. W. Luo, B. Liao, R. F. Li, W. Zhu, RNA secondary structure 3D graphical representation without degeneracy, *J. Math. Chem.* **39** (2006) 629–636.

[38] L. W. Liu, T. M. Wang, On 3D graphical representation of RNA secondary structures and their applications, *J. Math. Chem.* **42** (2007) 595–602.

[39] B. Liao, T. M. Wang, A 3D graphical representation of RNA secondary structures, *J. Biomol. Struct. Dyn.* **21** (2004) 827–832.

[40] B. Liao, J. W. Luo, R. F. Li, W. Zhu, RNA secondary structure 2D graphical representation without degeneracy, *Int. J. Quantum Chem.* **106** (2006) 1749–
–1755.

[41] J. F. Yu, X. Sun, J. H. Wang, A novel 2D graphical representation of protein sequence based on individual amino acid, *Int. J. Quantum Chem.* **111** (2011) 2835–2843.

[42] Y. H. Yao, Q. Dai, L. Li, X. Y. Nan, P. A. He, Y. Z. Zhang, Similarity/dissimilarity studies of protein sequences based on a new 2D graphical representation, *J. Comput. Chem.* **31** (2010) 1045–1052.

[43] M. Randić, M. Novič, A. R. Choudhury, D. Plavšić, On graphical representation of trans-membrane proteins, *SAR QSAR Environ. Res.* **23** (2012) 327–343.

[44] M. Randić, K. Mehulic, D. Vukičević, T. Pisanski, D. Vikić-Topić, D. Plavšić, Graphical representation of proteins as four-color maps and their numerical characterization, *J. Mol. Graph.* **27** (2009) 637–641.

[45] M. Randić, 2-D graphical representation of proteins based on physico-chemical properties of amino acids, *Chem. Phys. Lett.* **444** (2007) 176–180.

[46] B. Liao, B. Y. Liao, X. G. Lu, Z. Cao, A novel graphical representation of protein sequences and its application, *J. Comput. Chem.* **32** (2011) 2539–2544.

[47] C. Li, L. L. Xing, X. Wang, 2-D graphical representation of protein sequences and its application to coronavirus phylogeny, *BMB Rep.* **41** (2008) 217–222.

[48] P. He, A new graphical representation of similarity/dissimilarity studies of protein sequences, *SAR QSAR Environ. Res.* **21** (2010) 571–580.

[49] H. H. Bai, C. Li, H. Agula, J. Wang, L. L. Xing, CP-curve, a novel 3-D graphical representation of proteins, in: T. E. Simos, G. Maroulis (Eds.), *Computation in Modern Science and Engineering*, Amer. Inst. Phys., Melville, 2007, pp. 57–60.

[50] N. Liu, T. M. Wang, Graphical representations for protein secondary structure sequences and their application, *Chem. Phys. Lett.* **435** (2007) 127–131.

[51] M. Randić, M. Novič, M. Vračko, D. Plavšić, Study of proteome maps using partial ordering, *J. Theor. Biol.* **266** (2010) 21–28.

[52] M. Randić, Quantitative characterizations of proteome: Dependence on the number of proteins considered, *J. Proteome Res.* **5** (2006) 1575–1579.

[53] E. G. Giannopoulou, S. D. Garbis, A. Vlahou, S. Kossida, G. Lepouras, E. S. Manolakos, Proteomic feature maps: A new visualization approach in proteomics analysis, *J. Biomed. Inform.* **42** (2009) 644–653.

[54] S. C. Basak, B. D. Gute, Mathematical biodescriptors of proteomics maps: Background and applications, *Curr. Opin. Drug. Disc.* **11** (2008) 320–326.

[55] D. Bielinska-Wąż, T. Clark, P. Wąż, W. Nowak, A. Nandy, 2D-dynamic representation of DNA sequences, *Chem. Phys. Lett.* **442** (2007) 140–144.

[56] D. Bielinska-Wąż, W. Nowak, P. Wąż, A. Nandy, T. Clark, Distribution moments of 2D-graphs as descriptors of DNA sequences, *Chem. Phys. Lett.* **443** (2007) 408–413.

[57] D. Bielinska-Wąż, Graphical and numerical representations of DNA sequences: statistical aspects of similarity, *J. Math. Chem.* **49** (2011) 2345–2407.

[58] J. F. Yu, X. Sun, J. H. Wang, TN curve: A novel 3D graphical representation of DNA sequence based on trinucleotides and its applications, *J. Theor. Biol.* **261** (2009) 459–468.

[59] A. Ghosh, A. Nandy, Graphical representation and mathematical characterization of protein sequences and applications to viral proteins, *Adv. Protein Chem. Str.* **83** (2011) 1–42.

[60] A. Roy, C. Raychaudhury, A. Nandy, Novel techniques of graphical representation and analysis of DNA sequences - A review, *J. Biosci.* **23** (1998) 55–71.

[61] Y. H. Yao, Q. Dai, C. Li, P. A. He, X. Y. Nan, Y. Z. Zhang, Analysis of similarity/dissimilarity of protein sequences, *Proteins* **73** (2008) 864–871.

[62] H. J. Yu, D. S. Huang, Novel 20-D descriptors of protein sequences and it's applications in similarity analysis, *Chem. Phys. Lett.* **531** (2012) 261–266.

[63] Z. M. Nie, Z. F. Zhang, D. Wang, P. A. He, C. Y. Jiang, L. Song, F. Chen, J. Xu, L. Yang, L. L. Yu, J. Chen, Z. B. Lv, J. J. Lu, X. F. Wu, Y. Z. Zhang, Complete sequence and organization of Antheraea pernyi nucleopolyhedrovirus, a dr-rich baculovirus, *BMC Genomics* **8** (2007) #248.

[64] E. A. Herniou, J. A. Olszewski, D. R. O'Reilly, J. S. Cory, Ancient coevolution of baculoviruses and their insect hosts, *J. Virolo.* **78** (2004) 3244–3251.

[65] D. K. Thumbi, R. J. M. Eveleigh, C. J. Lucarotti, R. Lapointe, R. I. Graham, L. Pavlik, H. A. M. Lauzon, B. M. Arif, Complete sequence, analysis and organization of the orgyia leucostigma nucleopolyhedrovirus genome, *Viruses* **3** (2011) 2301–2327.

[66] Y. Jiang, F. Deng, H. L. Wang, Z. H. Hu, An extensive analysis on the global codon usage pattern of baculoviruses, *Arch. Virolo.* **153** (2008) 2273–2282.

[67] Y. Jiang, F. Deng, S. Rayner, H. L. Wang, Z. H. Hu, Evidence of a major role of GP64 in group I alphabaculovirus evolution, *Virus Res.* **142** (2009) 85–91.

[68] A. Nandy, P. Nandy, On the uniqueness of quantitative DNA difference descriptors in 2D graphical representation models, *Chem. Phys. Lett.* **368** (2003) 102–107.