BMB
reports

2-D graphical representation of protein sequences and its application to coronavirus phylogeny

Chun Li[1,*], *Lili Xing*[1] *& Xin Wang*[2]

[1]Department of Mathematics, Bohai University, Jinzhou, [2]Dalian Naval Academy, Dalian, P. R. China

Based on a five-letter model of the 20 amino acids, we propose a new 2-D graphical representation of protein sequence. Then we transform the 2-D graphical representation into a numerical characterization that will facilitate quantitative comparisons of protein sequences. As an application, we construct the phylogenetic tree of 56 coronavirus spike proteins. The resulting tree agrees well with the established taxonomic groups. [BMB reports 2008; 41(3): 217-222]

INTRODUCTION

The comparative study of DNA and proteins is a topic of considerable interest. For many years the standard procedure for comparison of different DNA sequences was based on computer-oriented and computer-intensive comparisons of sequences by sequence alignment. During the early 1990s, qualitative comparisons of DNA sequences were made possible by graphical representations of DNA, which allow visual inspection of lengthy sequences (1-24). It was later shown that it is possible to characterize numerically the graphical representation to obtain a numerical characterization of the degree of similarity /dissimilarity of different DNA sequences (9, 10, 14, 17-20). This is accomplished by associating with graphical representations of DNA a corresponding mathematical object such as a matrix, and then using various properties of mathematical object, like matrix invariants, as sequence descriptors. In this way one arrives at an alternative approach for comparative studies of DNA, which are less computer-intensive, because it replaces the original DNA sequence by an ordered set of sequence invariants, which can be viewed as components of vectors and thus comparison of sequences is transformed into a simpler comparison of vectors.

It should be mentioned that most of the graphical representations of DNA involve some degree of arbitrariness, such as the selection of directions to be assigned to individual bases. Therefore, extension of DNA graphical representations to those of proteins would increase enormously the number of possible alternative assignments for the 20 amino acids making such generalizations unacceptable, which is probably the most important reason why graphical representations of proteins have not been advanced. Recently, however, several schemes have been proposed in the literature which offered some progress for depicting proteins and thus allowing visual inspection of proteins (25-29). It is interesting that almost all of these representations of proteins are similar to the chaos game representation (CGR) of DNA sequences proposed by (6). In Jeffrey's CGR a DNA sequence is represented by a set of spots within a square. The four corners of a square are assigned labels A, T, G, and C, respectively. Graphical representation of DNA sequence is obtained by starting at the center of the square and moving half way from the center of the square to the corner corresponding to the first base of DNA. After that one continues to move from that spot half way towards the corner corresponding to the second base and so on till the whole DNA sequence is exhausted. Clearly, if one associates to each trio of nucleotides that define a triangle a single amino acid one can represent it by a spot placed in the center of the triangle. In this way Randić, M. constructed a 2-D graphical representation of the protein associated with DNA codons (25). Meanwhile Randić, M. *et al* replaced the two-dimensional representation of triplets based on the four corners of a square by an analogous three-dimensional representation based on assignment of triplets to the four corners of a tetrahedron, and then constructed a similar 3-D zigzag curve for protein sequences (26). The problem, however, is that for a given sequence of amino acids that defines a protein we often do not know the original DNA sequence. Without knowing the primary DNA sequence we cannot proceed to construct such graphs of the protein considered. The list of amino acids in a protein does not suffice to deduce what is the corresponding portion of the genetic code, because of the degeneracy of the genetic code. To solve this problem Randić, M. adopted a method of fictitious virtual genetic code, which assigns to each amino acid for which there are alternative codons a *single* codon (25). It is not difficult to see that the selected virtual genetic code is one of $6^3 \times 4^5 \times 3 \times 2^9$ possible such virtual codes in view of there being three amino acids associated with six codons, five with four codons, one with three and nine with two codons.

*Corresponding author. Tel: 86-0416-3400192; Fax: 86-0416-3400196; E-mail: lchlmb@yahoo.com.cn

Received 6 July 2007, Accepted 26 October 2007

Keywords: ALE-index, Coronavirus, Graphical representation, Phylogeny, Protein

In other words, the representative choices are to a degree arbitrary. In 2006, (28) proposed another 2-D graphical representation of proteins. Instead of Jeffrey's square with labels of 4 bases, they considered a unit circle, on the circumference of which at equal distances are positioned 20 amino acids. The graphical representation of proteins was obtained by starting in the center of the circle following amino acid sequence by moving half way towards to corresponding amino acid. In addition, by assigning the 20 amino acids to 20 points on the circumference of a unit circle (29) defined orientations for 20 branches of a star graph. Then they located vertices depicting multiple occurrence of each kind of amino acid on the corresponding branch. Consequently, a star-like graph of a protein sequence was obtained.

In this paper we will introduce a new 2-D graphical representation of proteins based on a five-letter model of the 20 amino acids. This approach is accompanied by a relatively small number of arbitrary choices associated with the graphical representation of proteins. Its application is shown by constructing the phylogenetic tree of 56 coronavirus spike proteins.

RESULTS AND DISCUSSION

Severe acute respiratory syndrome (SARS) is a newly emerged infectious disease that appeared in Guangdong Province, mainland China, in November 2002. By March 2003, the disease had spread globally and by July there were 8,447 probable SARS cases including 811 deaths reported from 32 countries or regions worldwide to the WHO (30-34). Although currently the spread of the virus seems to be confined to rigorous and timely quarantine measures, it may still be circulating in the animal reservoir and it is impossible to predict when it will return. Indeed, one should never be complacent when dealing with emerging infectious diseases.

Isolated in the mid-1960s, coronaviruses (order *Nidovirales*, family *Coronaviridae*, genus *Coronavirus*) are a diverse group of large, enveloped, positive-stranded RNA viruses that cause respiratory and enteric diseases in humans and other animals. Coronaviruses can be divided into three serologically distinct groups: two groups of predominantly mammalian coronaviruses, and a third group of avian coronaviruses (chicken and turkey). Phylogenetic analyses suggested that SARS-CoV may warrant assignment to a new, fourth Group within the genus *Coronavirus* (30, 31, 33-40). 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 the quantitative characterization of protein sequences, in what follows, we will construct the phylogenetic tree of 56 coronavirus spike proteins of Table 1.

As will be described later in more details a protein sequence can be associated with a 60-component vector. Suppose that there are two species i and j, the corresponding vectors are $v_i = (x_{i1}, x_{i2}, \cdots, x_{i,60})$ and $v_j = (x_{j1}, x_{j2}, \cdots, x_{j,60})$, respectively. Then we have

$$d_{ij} = \sqrt{\sum_{k=1}^{60}(x_{ik} - x_{jk})^2}$$

which denotes the distance between species i and j. Corresponding to 56 spike proteins, a 56×56 real symmetric matrix $D = (d_{ij})_{56\times 56}$ is obtained and used to reflect the evolutionary distance of the 56 coronavirus spike proteins. The phylogenetic tree (see Fig. 1) is constructed using the UPGMA program included in PHYLIP package v.3.65. The branch lengths are not scaled according to the distances and only the topology of the tree is concerned.

Observing Fig. 1, we find that SARS-CoVs appear to cluster together and form a separate branch, which can be distinguished easily from other three groups of coronaviruses. BCoV, BCoVM, BCoVQ, BCoVE, BCoVL, HCoV-OC43, MHV, MHVJHM, MHVA, MHVM and MHVP, which belong to group II, are situated at an independent branch. While PEDVC, PEDV, TGEVG and TGEV, belonging to group I, tend to cluster together. In another branch, the group III coronaviruses, including IBV, IBVC, IBVBJ, tend to cluster together. The resulting monophyletic clusters agree well with the established taxonomic groups. A closer look at the subtree of SARS-CoVs shows that SZ3 and SZ16, which belong to the animal epidemic phase, form a separate branch. Civet007, civet010, civet020, A022, B039, PC4-127, PC4-137, PC4-205 and GD03T0013, belonging to 03-04 interspecies epidemic, tend to cluster together, while all human SARS-CoVs of the 2003 epidemic tend to form another branch. It is noteworthy that the most recent SARS-CoV GD03T0013 (December 2003) is much closer to the palm civet SARS-like coronavirus than to any human SARS-CoV detected in the previous epidemic, which strengthens the argument for animal origin of the human SARS epidemic. In addition, the subtree of SARS-CoVs shows that the viruses from masked palm civets collected in 2003 are different from those of 2004 indicating the viral transmission from animal to human occurred independently in these two instances. This result is similar to that reported by other authors (41-43).

CONCLUSION

Based on a five-letter model of 20 amino acids, we first reduce a protein primary sequence into a five-letter sequence, which can be thought of as a coarse-grained description of the protein primary sequence. Although some information may be lost in the reduced sequences, we can focus our attention on the information of our interest. In particular, it makes the generalization from DNA graphical representations to those of proteins acceptable. In the next step, we give a 2-D graphical representation for the five-letter sequence, and then construct a 60-component vector, in which the normalized ALE-indices extracted from such 2-D graphs via L/L matrices are individual components, to characterize the protein primary sequence. Thus comparison of protein sequences is transformed into a

Table 1. The accession number, name and abbreviation for the 56 coronavirus spike proteins

NO.	Accession number	Name	Abbreviation
1	CAB91145	Transmissible gastroenteritis virus, genomic RNA	TGEVG
2	NP_058424	Transmissible gastroenteritis virus	TGEV
3	AAK38656	Porcine epidemic diarrhea virus strain CV777	PEDVC
4	NP_598310	Porcine epidemic diarrhea virus	PEDV
5	NP_937950	Human coronavirus OC43	HCoV-OC43
6	AAK83356	Bovine coronavirus isolate BCoV-ENT	BCoVE
7	AAL57308	Bovine coronavirus isolate BCoV-LUN	BCoVL
8	AAA66399	Bovine coronavirus strain Mebus	BCoVM
9	AAL40400	Bovine coronavirus strain Quebec	BCoVQ
10	NP_150077	Bovine coronavirus	BCoV
11	AAB86819	Mouse hepatitis virus strain MHV-A59C12 mutant	MHVA
12	YP_209233	Murine hepatitis virus strain JHM	MHVJHM
13	AAF69334	Mouse hepatitis virus strain Penn 97-1	MHVP
14	AAF69344	Mouse hepatitis virus strain ML-10	MHVM
15	NP_045300	Mouse hepatitis virus	MHV
16	AAP92675	Avain infectious bronchitis virus isolate BJ	IBVBJ
17	AAS00080	Avain infectious bronchitis virus strain Ca199	IBVC
18	NP_040831	Avain infectious bronchitis virus	IBV
19	AY304486	SARS coronavirus SZ3	SZ3
20	AY304488	SARS coronavirus SZ16	SZ16
21	AAS10463	SARS coronavirus GD03T0013	GD03T0013
22	AAU93318	SARS coronavirus PC4-127	PC4-127
23	AAV49720	SARS coronavirus PC4-137	PC4-137
24	AAU93319	SARS coronavirus PC4-205	PC4-205
25	AAU04646	SARS coronavirus civet007	civet007
26	AAU04649	SARS coronavirus civet010	civet010
27	AAU04664	SARS coronavirus civet020	civet020
28	AAV91631	SARS coronavirus A022	A022
29	AAV49730	SARS coronavirus B039	B039
30	AAP51227	SARS coronavirus GD01	GD01
31	AAS00003	SARS coronavirus GZ02	GZ02
32	AAP30030	SARS coronavirus BJ01	BJ01
33	AAP13567	SARS coronavirus CUHK-W1	CUHK-W1
34	AY394989	SARS coronavirus HZS2-D	HZS2-D
35	AY394992	SARS coronavirus HZS2-C	HZS2-C
36	AAP50485	SARS coronavirus FRA	FRA
37	AAP41037	SARS coronavirus TOR2	TOR2
38	AAQ01597	SARS coronavirus Taiwan TC1	TaiwanTC1
39	AAQ01609	SARS coronavirus Taiwan TC2	TaiwanTC2
40	AAP13441	SARS coronavirus Urbani	Urbani
41	AAQ94060	SARS coronavirus AS	AS
42	AAP30713	SARS coronavirus CUHK-Su10	CUHK-Su10
43	AAP33697	SARS coronavirus Frankfurt 1	Frankfurt1
44	AAP94737	SARS coronavirus CUHK-AG01	CUHK-AG01
45	AAP94748	SARS coronavirus CUHK-AG02	CUHK-AG02
46	AAP37017	SARS coronavirus TW1	TW1
47	AAR87523	SARS coronavirus TW2	TW2
48	BAC81348	SARS coronavirus TWH genomic RNA	TWH
49	BAC81362	SARS coronavirus TWJ genomic RNA	TWJ
50	AAP72986	SARS coronavirus HSR 1	HSR1
51	AAR23250	SARS coronavirus Sin01-11	Sino1-11
52	AAR23258	SARS coronavirus Sin03-11	Sino3-11
53	AY283794	SARS coronavirus Sin2500	Sin2500
54	AY283796	SARS coronavirus Sin2679	Sin2679
55	AAR14803	SARS coronavirus PUMC01	PUMC01
56	AAR14807	SARS coronavirus PUMC02	PUMC02

simpler comparison of vectors, which does not require multiple alignment.

SARS is a newly emerged infectious disease. Although SARS has been under control, it may still be circulating in the animal reservoir and it is impossible to predict when it will return. No effective drugs are currently available to cure this disease.

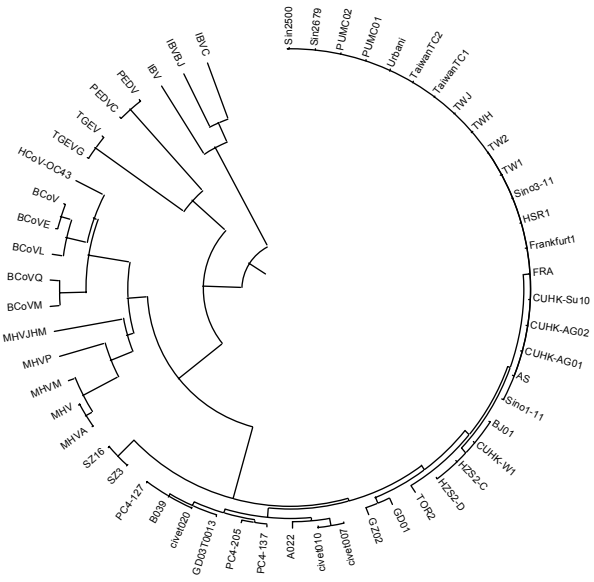

Fig. 1. The phylogenetic tree constructed with the current method.

Gaining insight into the phylogenetic relationships among coronaviruses would be helpful to discover drugs and develop vaccines against the SRAS-CoV. On the basis of the 60-component vectors, we infer phylogenetic relationship of 56 coronavirus spike proteins. The phylogenetic analysis shows that SARS-CoVs belong to a new cluster, named group IV. Furthermore, this group is divided into three subgroups, which correspond to the 03-animal epidemic, 03-04 interspecies epidemic, and human epidemic, respectively. These results are consistent with those of previous analyses.

MATERIALS AND METHODS

Dataset

Protein sequences of 56 coronaviruses used in this paper were downloaded from GenBank (see Table 1 for details).

The five-letter model of amino acids

Much effort has been made by considering minimalist models with a few types of amino acid residues to simplify the natural set of residues of 20 types for better physical understanding and practical purposes. In these models the compositions are much simpler than the real ones. The simplest reduction is the well-known HP model. The studies of such a model enable people to understand some fundamental physics and mechanism of protein folding. However, as argued in a number of studies (44, 45), the HP model may be too simple and lacks enough consideration on the heterogeneity and the complexity of the natural set of residues, such as the interactions between the residues. Moreover, the minimal sets of residues for protein design suggested by biochemical experiments seem unfavorable to those with only two types of residues since a small number of types obviously introduces the homopolymeric degeneracy. What is the suitable simplification for natural proteins, or how many types of residues are necessary for reproducing some useful structures and for a simplified representation of protein sequence characteristics? These are not well understood. Riddle, D. S. *et al* made an exciting approach to the problems mentioned above experimentally (46). By using combinatorial chemistry along with a screening strategy, they searched and found out a subset of the natural amino acids that can be used to construct a well-ordered proteinlike molecule consisting of β sheets. This subset contains five amino acids: isoleucine, alanine, glycine, glutamic acid and lysine, which are simply represented as I, A, G, E, and K. Three years after that, based on the statistical and the kinetic characteristics of the folding, and on the thermodynamic stability of the ground states of some reduced sequences, Wang and Wang proved that the suggested five-letter code is valid in general and feasible for elucidating characteristics of real proteins with 20 kinds of amino acids (44, 45). We will in the next subsection describe a particular construction of graphical representation of proteins based on the five-letter model of 20 amino acids.

Construction of the 2-D graphical representation

Following the method introduced by Wang *et al*, the 20 amino acids can be classified into five groups (44, 45):

group-I = {C, M, F, I, L, V, W, Y},
group-II = {A, T, H},
group-III = {G , P},
group-IV = {D , E},
group-V = {S, N, Q, R, K}.

Each group contains some residues which interact with others in a similar way. Moreover, the five groups have representative residues, they are: I, A, G, E, and K, respectively. Thus a protein primary sequence can be reduced into a five-letter sequence by substituting (or replacing) each letter with its representative letter. For example, the five-letter sequence of MRSLIYFWLLLPVLPTLSLPQDVTRCQSTT, the first 30 amino acid residues of the spike protein sequence of PEDV, is IKKIIIIIIIIGIIGAIKIGKEIAKIKKAA. The five-letter sequence may be regarded as a coarse-grained description of the protein primary sequence. Via comparisons of the reduced sequences it will be easier to understand the biological function of various kinds of amino acid residues.

For a given five-letter sequence, similar to data previously done for DNA sequence, we draw five horizontal lines separated by unit distances, on which dots representing the letter constituting the considered sequence are placed (17). The representation requires first to associate the five letters with the five horizontal lines. We assign G to the first line, I to the second, K to the third, A to the fourth and E to the last. The consecutive letters are placed along the horizontal axes at unit distance intervals. In the next step we connect adjacent dots with lines, which results in a zigzag like curve of a definite geometrical form (Fig. 2).

Fig. 2. The 2-D graphical representation of the five-letter sequence IKKIIII-IIIIGIIGAIKIGKEIAKIKKAA.

Fig. 3. The curve of the five-letter sequence IKKIIIIIIIIGIIGAIKIGKEIAKIKKAA.

Numerical characterization of proteins

In this section, we give a numerical characterization of the 2-D graphical representation that will facilitate quantitative comparisons of protein sequences. One of the possibilities to achieve this aim is to characterize the graphs by invariants. To do so, we transform the graphical representation into another mathematical object, a matrix. The matrices associated with a graph include ED, D/D, L/L, and their 'higher order' matrices (9, 10, 14, 17-20, 26, 47-49). Among them, L/L is a matrix whose elements are defined as the quotient of the Euclidean distance between a pair of vertices (dots) of the zigzag curve and the sum of distances between the same pair of vertices measured along the zigzag curve. Once a real symmetric matrix M is given, one often uses some of matrix invariants, such as the average matrix element, the average row sum, the leading eigenvalue, and the Wiener number, as descriptors of the sequence (9, 14, 17-20, 26, 47-49). Moreover, in our previous paper (48), an alternative sequence invariant called 'ALE-index' was proposed. The ALE-index is defined as

$$\chi = \chi(M) = \frac{1}{2}\left(\frac{1}{n}\|M\|_{m1} + \sqrt{\frac{n-1}{n}}\|M\|_F\right),$$

where $\|M\|_{m1} \equiv \sum_{i,j=1}^{n} |a_{ij}|$, $\|M\|_F \equiv \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^2} = \sqrt{tr(M^T M)}$.

Clearly, the ALE-index is very simple for calculation so that it can be directly used to handle long sequences. If desired, one can introduce weighting procedure that will normalize magnitudes of the ALE-indices to reduce variations caused by comparison of matrices of different size. For instance, one can consider instead of χ a normalized ALE-index $\chi' = \chi / n$, where n is the length of the sequence and the order of the corresponding matrix as well. For example, the normalized ALE-index of the L/L matrix corresponding to the curve of Fig. 2 is easily calculated as 0.6944.

In Fig.2, we get a curve by assigning the five letters to the five horizontal lines in the order G-I-K-A-E. If we assign them in order I-A-G-E-K, we can obtain another curve (Fig. 3). It is not difficult to see that the labels I, A, G, E and K can be arranged in 5 ways. Since for a given order of labels for the five horizontal lines the reverse order of labels does not introduce a novelty, there are at most 60 essentially different patterns of the zigzag curves representing the same five-letter sequence. Therefore, a protein primary sequence can be characterized by a 60-component vector in which the normalized ALE-indices of the corresponding L/L matrices are individual components.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China and the Science Research Project of Educational Department of Liaoning Province.

REFERENCES

1. Bielińska-Wąż, D., Clark, T., Wąż, P., Nowak, W. and Nandy, A. (2007) 2D-dynamic representation of DNA sequences. *Chem. Phys. Lett.* **442,** 140-144.
2. Bielińska-Wąż, D., Nowak, W., Wąż, P., Nandy, A. and Clark, T. (2007) Distribution moments of 2D-graphs as descriptors of DNA sequences. *Chem. Phys. Lett.* **443,** 408-413.
3. Gates, M. A. (1986) A simple way to look at DNA. *J. Theor. Biol.* **119,** 319-328.
4. Guo, X. F., Randić, M. and Basak, S. C. (2001) A novel 2-D graphical representation of DNA sequences of low degeneracy. *Chem. Phys. Lett.* **350,** 106-112.
5. Hamori, E. and Ruskin, J. (1983) H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *J. Biol. Chem.* **258,** 1318.
6. Jeffrey, H. I. (1990) Chaos game representation of gene structure. *Nucleic Acid Res.* **18,** 2163-2170.
7. Leong, P. M. and Morgenthaler, S. (1995) Random walk and gap plots of DNA sequences. *Comput. Applic. Biosci.* **12,** 503-511.
8. Li, C. and Wang, J. (2004) On a 3-D representation of DNA primary sequences. *Comb. Chem. High T. Scr.* **7,** 23-27.
9. Li, C., Tang, N. N. and Wang, J. (2006) Directed graphs of DNA sequences and their numerical characterization. *J. Theor. Biol.* **241,** 173-177.
10. Li, C. and Hu, J. (2006) 2-D Graphical representation for characteristic sequences of DNA and its application. *J. Biochem. Mol. Biol.* **39,** 292-296.

11. Nandy, A. (1994) A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. *Curr. Sci.* **66**, 309-313.
12. Nandy, A. (1994) Graphical representation of long DNA sequences. *Curr. Sci.* **66**, 821.
13. Nandy, A., Harle, M. and Basak, S. C. (2006) Mathematical descriptors of DNA sequences: development and applications. *ARKIVOC* **9**, 211-238.
14. Randić, M., Vracko, M., Nandy, A. and Basak, S. C. (2000) On 3-D graphical representation of DNA primary sequence and their numerical characterization. *J. Chem. Inf. Comput. Sci.* **40**, 1235-1244.
15. Randić, M., Guo, X. F. and Basak S. C. (2001) On the Characterization of DNA primary sequences by triplet of nucleic acid bases. *J. Chem. Inf. Comptu. Sci.* **41,** 619-626.
16. Randić, M. and Balaban, A. T. (2003) On a four-dimensional representation of DNA primary sequences. *J. Chem. Inf. Comptu. Sci.* **43,** 532-539.
17. Randić, M., Vracko, M., Lers, N. and Plavsić, D. (2003) Novel 2-D graphical representation of DNA sequences and their numerical characterization. *Chem. Phys. Lett.* **368,** 1-6.
18. Randić, M., Vracko, M., Lers, N. and Plavsić, D. (2003) Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. *Chem. Phys. Lett.* **371,** 202-207.
19. Randić, M., Vracko, M., Zupan, J. and Novic M. (2003) Compact 2-D graphical representation of DNA. *Chem. Phys. Lett.* **373,** 558-562.
20. Randić, M. (2004) Graphical representations of DNA as 2-D map. *Chem. Phys. Lett.* **386,** 468-471.
21. Randić, M. and Zupan, J. (2004) Highly compact 2-D graphical representation of DNA sequences. *SAR QSAR Environ. Res.* **15,** 191-205.
22. Roy, A., Raychaudhury, C. and Nandy, A. (1998) A novel technique of graphical representation and analysis of DNA sequences-A review. *J. Biosci.* **23,** 55-71.
23. Wu, Y. H., Liew, A. W., Yan, H. and Yang, M. (2003) DB-Curve: a novel 2D method of DNA sequence visualization and representation. *Chem. Phys. Lett.* **367,** 170-176.
24. Zhang, R. and Zhang, C. T. (1994) Z curves, an intuitive tool for visualizing and analyzing DNA sequences. *J. Biomol. Struc. Dyn.* **11,** 767-782.
25. Randić, M. (2004) 2-D Graphical representation of proteins based on virtual genetic code. *SAR QSAR Environ. Res.* **15,** 147-157.
26. Randić, M., Zupan, J. and Balaban, A. T. (2004) Unique graphical representation of protein sequences based on nucleotide triplet codons. *Chem. Phys. Lett.* **397,** 247-252.
27. Randić, M., Balaban, A. T., Novic, M., Zaloznik, A. and Pisanski, T. (2005) A novel graphical representation of proteins. *Period. Boil.* **107,** 403-414.
28. Randić, M., Butina, D. and Zupan, J. (2006) Novel 2-D graphical representation of proteins. *Chem. Phys. Lett.* **419,** 528-532.
29. Randić, M., Zupan, J. and Vikić-Topić, D. (2007) On representation of proteins by star-like graphs. *J. Mol. Graph. Model.* **26,** 290-305.
30. Lau, S. K. P., Wo, P. C. Y. and Li, K. S. M., *et al*. (2005) Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *PNAS* **102,** 14040-14045.
31. Marra, M. A., Jones, S. J. M. and Astell, C. R., *et al*. (2003) The genome sequence of the sars-associated coronavirus. *Science* **300,** 1399.
32. Poon, L. L., Chu, D. K., Chan, K. H., Wong, O. K., Ellis T. M., Leung, Y. H., Lau, S. K., Woo, P. C., Suen, K. Y., Yuen, K. Y., Guan, Y. and Peiris, J. S. (2005) Identification of a novel coronavirus in bats. *J. Virol.* **79,** 2001-2009.
33. Rota, P. A., Oberste, M. S. and Monroe, S. S., *et al*. (2003) Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* **300,** 1394.
34. Satija, N. and Lal, S. (2007) The Molecular Biology of SARS Coronavirus. *Ann. N.Y. Acad. Sci.* **1102,** 26-38.
35. Gao, L., Qi, J., Wei, H. B., Sun, Y. G. and Hao, B. L. (2003) Molecular phylogeny of coronaviruses including human SARS-CoV. *Chin. Sci. Bull.* **48,** 1170-1174.
36. Gorbalenya, A. E., Snijder, E. J. and Spaan, W. J. M. (2004) Severe acute respiratory syndrome coronavirus phylogeny: toward consensus. *J. Virol.* **78,** 7863-7866.
37. Ksiazek, T. G., Zaki, S. R. and Urbani, C., *et al*. (2003) A novel coronavirus associated with severe acute respiratory syndrome. *N. Engl. J. Med.* **348,** 1953-1966.
38. Skowronski, D. M., Astell, C., Brunham, R. C., Low, D. E., Petric, M., Roper, R.L., Talbot, P. J., Tam, T. and Babiuk, L. (2005) Severe acute respiratory syndrome (SARS): a year in review. *Annu. Rev. Med.* **56,** 357-381.
39. Snijder, E. J., Bredenbeek, P. J. and Dobbe, J. C., *et al*. (2003) Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *J. Mol. Biol.* **331,** 991-1004.
40. Zheng, W. X., Chen, L. L., Ou, H. Y., Gao, F. and Zhang, C. T. (2005) Coronavirus phylogeny based on a geometric approach. *Mol. Phylogenet. Evol.* **36,** 224-232.
41. Chinese SARS Molecular Epidemiology Consortium. (2004) Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science* **303,** 1666-1669.
42. Shi, Z. and Hu, Z. (2007) A review of studies on animal reservoirs of the SARS coronavirus. *Virus Res.* in press.
43. Song, H. D., Tu, C. C. and Zhang, G. W., *et al*. (2005) Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. *PNAS* **102,** 2430-2435.
44. Wang, J. and Wang, W. (1999) A computational approach to simplifying the protein folding problem. *Nat. Struct. Biol.* **6,** 1033 -1038.
45. Wang, J. and Wang, W. (2000) Modeling study on the validity of a possibly simplified representation of proteins. *Phys. Rev. E.* **61,** 6981-6986.
46. Riddle, D. S., Santiago, J. V., Brayhall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q. and Baker, D. (1997) Functional rapidly folding proteins from simplified amino acid sequences. *Nat. Struct. Biol.* **4,** 805-809.
47. Jaklic, G., Pisanski, T. and Randić, M. (2006) Characterization of Complex Biological Systems by Matrix Invariants. *J. Comput. Biol.* **13,** 1558-1564.
48. Li, C. and Wang, J. (2005) New Invariant of DNA Sequences. *J. Chem. Inf. Model.* **45,** 115-120.
49. Randić, M., Zupan, J., Novic, M., Gute, B. D. and Basak, S. C. (2002) Novel matrix invariants for characterization of changes of proteomics maps. *SAR QSAR Environ. Res.* **13,** 689-703.