Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology)
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

New Technique:

Protein sequence analysis based on hydropathy profile of amino acids[*]

Xiao-li XIE[†1,2], Li-fei ZHENG[1], Ying YU[3], Li-ping LIANG[1], Man-cai GUO[1], John SONG[4], Zhi-fa YUAN[†‡1]

(*[1]College of Sciences, Northwest A&F University, Yangling 712100, China*)

(*[2]College of Animal Science and Technology, Northwest A&F University, Yangling 712100, China*)

(*[3]College of Animal Science and Technology, China Agriculture University, Beijing 100193, China*)

(*[4]Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA*)

[†]E-mail: xiemary@nwsuaf.edu.cn; zhifayuan@nwsuaf.edu.cn

Received Feb. 23, 2011; Revision accepted July 21, 2011; Crosschecked Nov. 16, 2011

Abstract: Biology sequence comparison is a fundamental task in computational biology. According to the hydropathy profile of amino acids, a protein sequence is taken as a string with three letters. Three curves of the new protein sequence were defined to describe the protein sequence. A new method to analyze the similarity/dissimilarity of protein sequence was proposed based on the conditional probability of the protein sequence. Finally, the protein sequences of ND6 (NADH dehydrogenase subunit 6) protein of eight species were taken as an example to illustrate the new approach. The results demonstrated that the method is convenient and efficient.

Key words: Protein sequence, Sequence comparison, Similarity/dissimilarity, Conditional probability
doi:10.1631/jzus.B1100052 **Document code:** A **CLC number:** Q811.4

1 Introduction

The comparative biological sequence is one of the issues in bioinformatics when analyzing similarities of function and properties of different sequences. Similarly, evolutionary homology is analyzed by comparing DNA and protein sequences. In general, there are two types of methodologies to conduct the comparison. One is an alignment-based method, and the other is an alignment-free method.

Sequence alignment is based on computer-oriented and computer-intensive comparisons of sequences, and then a distance function or a score function is obtained. Using the distance function, one can compare biological sequences. However, multiple sequence alignment of several hundred sequences always produces a bottleneck, firstly due to long computational time, and secondly due to possible bias of multiple sequence alignments for multiple occurrences of highly similar sequences (Pham and Zuegg, 2004). Therefore, the emergence of a study on alignment-free sequence analysis is obvious. Until now, alignment-free sequence analysis is still in its early development. For most alignment-free methods, a biological sequence should be transformed into an object for which a linear algebra and statistical theory already has useful analytical tools. Since 1983, DNA sequence has been represented in different dimension spaces (Hamori and Ruskin, 1983; Hamori, 1985; Nandy, 1994; 1996; Nandy and Basak, 2000; Randić *et al.*, 2001; Randić, 2003; Randić and Balaban, 2003; Zhang *et al.*, 2003; Liao and Wang, 2004; Liao *et al.*, 2005; Nandy *et al.*, 2006; Bai *et al.*, 2007; Feng and Wang, 2008). Each nucleotide of a given DNA sequence is a point in different dimension spaces, and

[‡] Corresponding author

[*] Project (No. Z111020834) supported by 08 Special Talent Fund of Northwest A&F University, China

these graphical representations can allow us to qualitatively analyze DNA sequences, and provide a way of viewing, sorting and comparing various genomic sequences. Based on the graphical representation, it is possible to numerically characterize DNA sequence and further quantitatively measure similarity of different DNA sequences. Although protein sequence and DNA sequence belong to symbolic sequences, compared with DNA sequence, there are fewer methods for the graphical representation of protein sequence. This is mainly because extension of DNA graphical representation to protein sequences would enormously increase the number of possible alternative assignments for the 20 amino acids. The amino acid sequence is the key to understanding protein structure and function in the cell, so analysis of amino acid sequence is an important part of post-genomic studies. Recently, several schemes have been proposed in protein graphical representation (Randić and Krilov, 1997; Vinga and Almeida, 2003; Bai and Wang, 2005; Li J. *et al*., 2006; Li C. *et al*., 2008; Munteanu *et al*., 2008; Yau *et al*., 2008; Yao *et al*., 2008; 2009; Wen and Zhang, 2009). In order to plot amino acid sequence, 20 amino acids in protein sequences are divided into different types, including protein sequence regarded as a word with three, four, or five different letters. Since ordering amino acids based on their physicochemical properties may offer better insights into comparative study of protein than representation of protein based on the random ordering of amino acid, Randić (2007) and Yao *et al.* (2008; 2009) outlined different 2D graphical representations of protein sequence based on different physicochemical properties. The graphical representation of protein sequence cannot only describe amino acid sequence, but also measure similarity/dissimilarity of different protein sequences. However, the methods only consider the string's information of protein, and do not consider adjacent string's information of amino acid sequence. Here, we choose conditional probability to measure adjacent string's information.

In this paper, we converted a protein sequence into three-letter sequence based on hydropathy profile of amino acid and defined the three curves to represent different hydropathy features. We then selected conditional probability as a new invariant for the protein sequences. To illustrate the proposed method, we made a comparison of the sequences belonging to eight ND6 (NADH dehydrogenase subunit 6) proteins from http://www.ncbi.nlm.nih.gov/: *human* (YP_003024037.1), *gorilla* (NP_008223), *chimpanzee* (NP_008197), *wallaroo* (NP_007405), *harbor seal* (*H. seal*) (NP_006939), *gray seal* (*G. seal*) (NP_007080), *rat* (AP_004903), and *mouse* (NP_904339).

2 Protein feature sequence

According to the hydropathy profile of amino acids, the amino acids can be classified into three groups (Nei and Kumar, 2002; Liu and Wang, 2006): internal group (F, I, L, M, V), external group (D, E, H, K, N, Q, R), and ambivalent group (S, T, Y, C, W, G, P, A). The amino acid of internal group tends to occur in the inner side of the protein's spatial structure, while the amino acid of external group tends to appear at the surface. In order to characterize the hydropathicity of a protein primary structure, we defined a primary protein sequence as a symbolic sequence including three letters according to the following rule:

$$F(S(i))=\begin{cases}\mathrm{I} & S(i)=\mathrm{F, I, L, M, V},\\ \mathrm{E} & S(i)=\mathrm{D, E, H, K, N, Q, R},\\ \mathrm{A} & S(i)=\mathrm{S, T, Y, C, W, G, P, A},\end{cases}\qquad(1)$$

where $S(i)$ is the letter in the ith position in the protein primary sequence, and $F(S(i))$ is the substitution for $S(i)$. Since the hydropathy profile can detect more evolutionary relationships, in the next section, we analyzed the new protein sequence containing three letters through different mathematical methods.

3 Graphical representation of protein sequence

Given a protein primary sequence with length N, we transformed it into a new sequence according to the above definition. For example, for the protein sequence, S=MMYALFLLSVGLVMGFVGFS, then $F(S)$=IIAAIIIIAIAIIIAIIAIA. To obtain more information, we defined three curves of the sequence. Firstly, we let

$$X_i^{\mathrm{IE}}=\begin{cases}+1 & \text{if } F(S(i))=\mathrm{I},\\ 0 & \text{otherwise},\\ -1 & \text{if } F(S(i))=\mathrm{E};\end{cases}$$
$$X_i^{\mathrm{EA}}=\begin{cases}+1 & \text{if } F(S(i))=\mathrm{E},\\ 0 & \text{otherwise},\\ -1 & \text{if } F(S(i))=\mathrm{A};\end{cases}\qquad (2)$$
$$X_i^{\mathrm{IA}}=\begin{cases}+1 & \text{if } F(S(i))=\mathrm{I},\\ 0 & \text{otherwise},\\ -1 & \text{if } F(S(i))=\mathrm{A};\end{cases}$$

where i ranges from 1 to N. Then, let

$$Y_0=0,\quad Y_n^u=Y_0+\sum_{i=1}^{n}X_i^u\quad (u=\mathrm{IE},\mathrm{IA},\mathrm{EA}).\qquad (3)$$

Y_n^u and n are Y axis and X axis, respectively, and then we can draw three different curves, which are named as IE, IA, and EA curves of the protein sequence. The three different curves can give us some information about the protein sequence. According to the IE curve, we can compare the numbers of the amino acids belonging to the internal group and the external group at different positions. The IA curve can then be used to compare the numbers of the amino acids belonging to the internal group and the ambivalent group at different positions. Finally, the EA curve can compare the numbers of the amino acids of the external group and the ambivalent group at different positions. According to the above definitions of three different curves, we drew three curves of ND6 proteins for the eight species (Fig. 1).

Fig. 1 shows that the amino acids of the internal group in ND6 protein sequences are more than the amino acids of the external group, and the amino acids of the ambivalent group are more than the amino acids of the external group. Furthermore, it is evident that *G. seal* and *H. seal* have similar curves, *rat* and *mouse*'s curves are almost identical, and the three curves of *human*, *gorilla*, and *chimpanzee* are similar, but *wallaroo*'s curve is different from curves of other species.

4 Numerical characterizations of protein sequences

Protein sequence is composed of three parts, internal group, external group and ambivalent group, so we regard the random numerical sequence to be composed of three parts (+1, 0, −1). We calculated the conditional probability, which was invariant to quantity protein sequences.

For example, let X_i^{IE} represents the state of the ith (i=1, 2, ..., N) moment, state space S={+1, 0, −1}. There are nine conditional probabilities as follows:

$$\begin{cases}p(\mathrm{A}|\mathrm{A})=p_{0,0}^{\mathrm{IE}}=p(x_{i+1}=0|x_i=0),\\ p(\mathrm{A}|\mathrm{I})=p_{0,1}^{\mathrm{IE}}=p(x_{i+1}=0|x_i=1),\\ p(\mathrm{A}|\mathrm{E})=p_{0,-1}^{\mathrm{IE}}=p(x_{i+1}=0|x_i=-1),\\ p(\mathrm{I}|\mathrm{A})=p_{1,0}^{\mathrm{IE}}=p(x_{i+1}=1|x_i=0),\\ p(\mathrm{I}|\mathrm{I})=p_{1,1}^{\mathrm{IE}}=p(x_{i+1}=1|x_i=1),\\ p(\mathrm{I}|\mathrm{E})=p_{1,-1}^{\mathrm{IE}}=p(x_{i+1}=1|x_i=-1),\\ p(\mathrm{E}|\mathrm{A})=p_{-1,0}^{\mathrm{IE}}=p(x_{i+1}=-1|x_i=0),\\ p(\mathrm{E}|\mathrm{I})=p_{-1,1}^{\mathrm{IE}}=p(x_{i+1}=-1|x_i=1),\\ p(\mathrm{E}|\mathrm{E})=p_{-1,-1}^{\mathrm{IE}}=p(x_{i+1}=-1|x_i=-1).\end{cases}\qquad (4)$$

According to the above definition, we can obtain these conditional probabilities of a given protein sequence. The conditional probability of each of ND6 proteins is listed in Table 1.

Table 1 Conditional probabilities of amino acids of eight species

Species	p(I|I)	p(E|I)	p(A|I)	p(I|E)	p(E|E)	p(A|E)	p(I|A)	p(E|A)	p(A|A)
Human	0.5375	0.1125	0.3500	0.2727	0.1818	0.5455	0.4306	0.1250	0.4444
Gorilla	0.5385	0.1026	0.3590	0.2727	0.1818	0.5455	0.4054	0.1351	0.4595
Chimpanzee	0.5325	0.1169	0.3507	0.2381	0.1429	0.6191	0.4079	0.1184	0.4737
Wallaroo	0.5500	0.1125	0.3375	0.4737	0.1579	0.3684	0.3971	0.1029	0.5000
H. seal	0.5556	0.0741	0.3704	0.3182	0.2727	0.4091	0.4028	0.1389	0.4583
G. seal	0.5625	0.0625	0.3750	0.2727	0.2727	0.4546	0.3973	0.1507	0.4521
Mouse	0.5488	0.1098	0.3415	0.3333	0.2381	0.4286	0.4348	0.1015	0.4638
Rat	0.5366	0.1098	0.3537	0.3810	0.2381	0.3810	0.4348	0.1015	0.4638

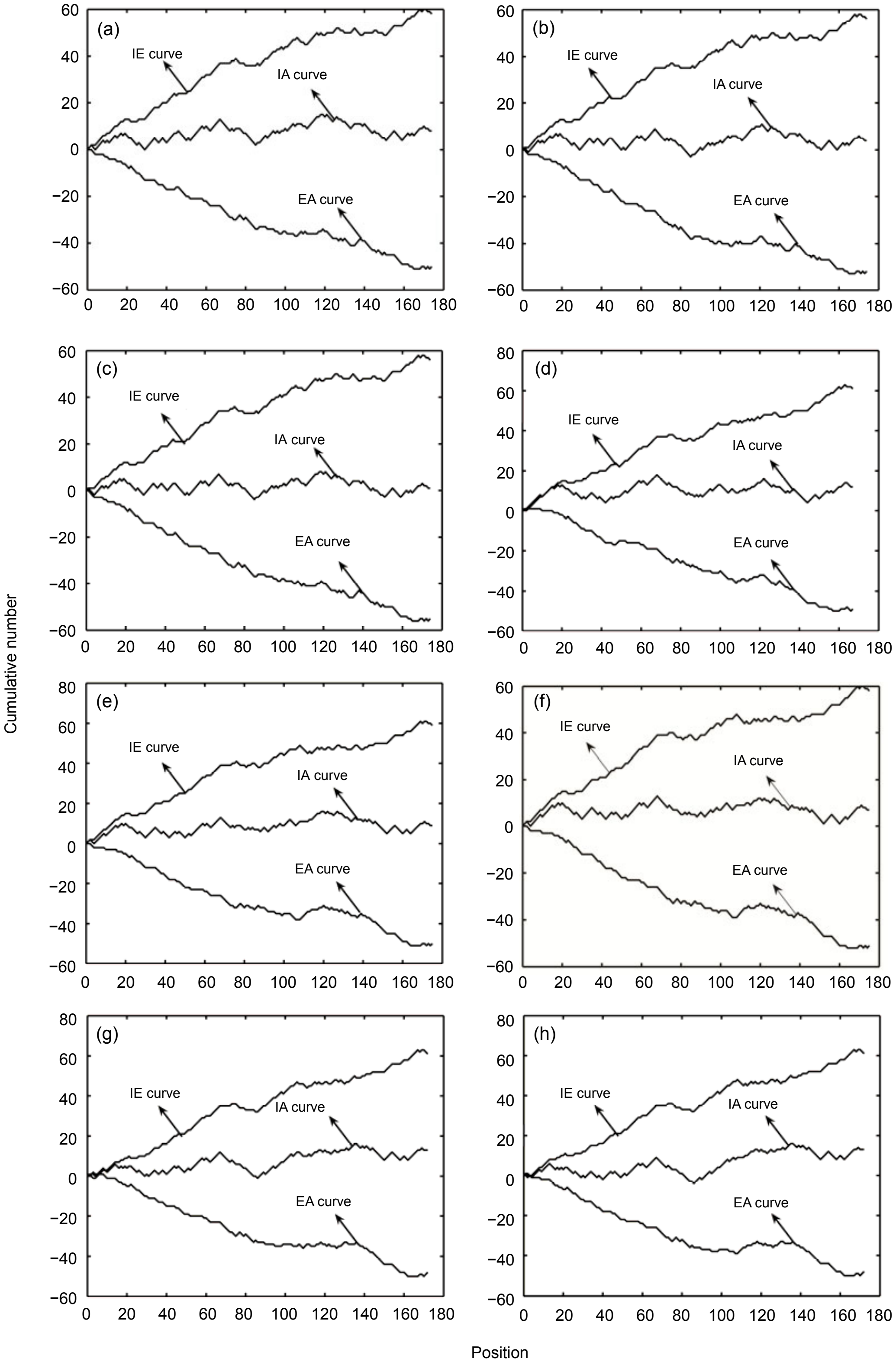

Fig. 1 Three curves of ND6 proteins

(a) *Human*; (b) *Gorilla*; (c) *Chimpanzee*; (d) *Wallaroo*; (e) *H. seal*; (f) *G. seal*; (g) *Mouse*; (h) *Rat*

5 Similarity/dissimilarity analysis

Given two protein sequences, we can obtain two nine-component vectors whose elements are conditional probabilities for each protein sequence. Based on the vectors, we can compare different protein sequences. In general, similarities of the two vectors can be obtained by calculating Euclidean distance. The smaller the Euclidean distance of two vectors is, the more similar are the protein sequences. The Euclidean distance of two vectors ***u*** and ***v*** is as follows:

$$\mathrm{d}(\boldsymbol{u},\boldsymbol{v})=\sqrt{\sum_{i=1}^{k}(u_i-v_i)^2}, \tag{5}$$

where u_i and v_i denote the components of vectors ***u*** and ***v***, respectively. k is the dimension of vectors ***u*** and ***v***. Yao *et al*. (2009) proposed a new similarity measure of sequences, and coefficient of determination (r^2), which is defined as:

$$r^2=\frac{\left(k\sum_{i=1}^{k}u_iv_i-\sum_{i=1}^{k}u_i\sum_{i=1}^{k}v_i\right)^2}{\left(k\sum_{i=1}^{k}u_i^2-(\sum_{i=1}^{k}u_i)^2\right)\left(k\sum_{i=1}^{k}v_i^2-(\sum_{i=1}^{k}v_i)^2\right)}. \tag{6}$$

r^2 can vary from 0 to 1, and represents the percent of the data, which is the closest to the line of best fit. The larger the coefficient of determination of two vectors is, the more similar are the protein sequences. In Tables 2 and 3, we give the similarity/dissimilarity matrices for the eight ND6 sequences based on Euclidean distance and coefficient of determination amongst nine-component vectors. As shown in Tables 2 and 3, it is obvious that ND6 proteins of *human*, *gorilla*, and *chimpanzee* are more similar to each other. In addition, ND6 proteins are more similar for (*G. seal*, *H. seal*) and (*mouse*, *rat*). However, ND6 protein of *wallaroo* is very dissimilar to others amongst the eight species. The results are consistent with the known fact of evolution (Yao *et al*., 2009).

6 Discussion and conclusions

Biology sequence analysis is a fundamental task in computational biology, whose aim is to detect similarity/dissimilarity relationships between molecular sequences. Some alignment-free methods to analyze similarities/dissimilarities of DNA sequences have been proposed. However, there are few alignment-free methods to analyze protein sequences. The amino

Table 2 Similarity/dissimilarity matrices based on the Euclidean distances among the nine-component vectors consisting of conditional probabilities

Species	*Gorilla*	*Chimpanzee*	*Wallaroo*	*H. seal*	*G. seal*	*Mouse*	*Rat*
Human	0.0338	0.0979	0.2780	0.1797	0.1487	0.1472	0.2071
Gorilla		0.0945	0.2752	0.1737	0.1390	0.1516	0.2099
Chimpanzee			0.3465	0.2661	0.2263	0.2365	0.2956
Wallaroo				0.2114	0.2639	0.1803	0.1356
H. seal					0.0674	0.0801	0.1015
G. seal						0.1143	0.1602
Mouse							0.0695

Table 3 Similarity/dissimilarity matrices based on coefficients of determination among the nine-component vectors consisting of conditional probabilities

Species	*Gorilla*	*Chimpanzee*	*Wallaroo*	*H. seal*	*G. seal*	*Mouse*	*Rat*
Human	0.9950	0.9753	0.6920	0.8585	0.9030	0.9004	0.8114
Gorilla		0.9777	0.6980	0.8684	0.9155	0.8926	0.8069
Chimpanzee			0.5959	0.7557	0.8253	0.8046	0.6944
Wallaroo				0.8078	0.7082	0.8582	0.9254
H. seal					0.9782	0.9560	0.9454
G. seal						0.9236	0.8732
Mouse							0.9687

acid sequence of a protein is the key to understanding its structure and function in the cell, so we present a new method to analyze protein primary sequence in this paper.

The method is based on the graphical representation and conditional probability taken as the numerical characterization of the protein sequence. The demonstrable significance of the new method is that it cannot only analyze similarity/dissimilarity of protein sequences, but also provide more biological information about the protein sequences. According to the IE curve, we can compare the numbers of amino acids of the internal and external groups at different positions. Also the IA curve can be used to compare the numbers of amino acids of the internal and ambivalent groups at different positions. The EA curve can be used to compare the numbers of amino acids in the external and ambivalent groups at different positions. Therefore the three curves show the distribution of the three types of amino acids. Furthermore, the conditional probability reflected the distribution of the two adjacent amino acids. The new approach was applied to ND6 protein sequences of several species and results have shown that the introduction of hydropathy profile of amino acids into protein sequence is effectual and feasible.

Acknowledgements

We would like to thank Dr. Jian-gang WANG (College of Animal Science and Technology, Northwest A&F University, China), Jian-zhong LUO (Department of Foreign languages, Northwest A&F University, China), Feng AN and Dr. Jun-li DU (College of Sciences, Northwest A&F University, China) for their helpful suggestions.

References

Bai, F., Wang, T., 2005. A 2-D graphical representation of protein sequences based on nucleotide triplet codons. *Chem. Phys. Lett.*, **413**(4-6):458-462. [doi:10.1016/j.cplett.2005.08.011]

Bai, F., Liu, Y., Wang, T., 2007. A representation of DNA primary sequences by random walk. *Math. Biosci.*, **209**(1):282-291. [doi:10.1016/j.mbs.2006.06.004]

Feng, J., Wang, T., 2008. A 3D graphical representation of RNA secondary structures based on chaos game representation. *Chem. Phys. Lett.*, **454**(4-6):355-361. [doi:10.1016/j.cplett.2008.01.041]

Hamori, E., 1985. Novel DNA sequence representation. *Nature*, **314**(6012):585-586. [doi:10.1038/314585a0]

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

Li, C., Xing, L., Wang, X., 2008. 2-D graphical representation of protein sequences and its application to coronavirus phylogeny. *BMB Rep.*, **41**(3):217-222. [doi:10.5483/BMBRep.2008.41.3.217]

Li, J., Li, F., Wang, W., 2006. Simplification of protein sequence and alignment-free sequence analysis. *Prog. Biochem. Biophys.*, **33**(12):1215-1222 (in Chinese).

Liao, B., Wang, T., 2004. Analysis of similarity of DNA sequences based on 3D graphical representation. *Chem. Phys. Lett.*, **388**(1-3):195-200. [doi:10.1016/j.cplett.2004.02.089]

Liao, B., Tan, M., Ding, K., 2005. Application of 2D graphical representation of DNA sequence. *Chem. Phys. Lett.*, **414**(4-6):296-300. [doi:10.1016/j.cplett.2005.08.079]

Liu, N., Wang, T., 2006. Protein-based phylogenetic analysis by using hydropathy profile of amino acids. *FEBS Lett.*, **580**(22):5321-5327. [doi:10.1016/j.febslet.2006.08.086]

Munteanu, C.B., Gonzalez-Diaz, H., Magalhaes, A.L., 2008. Enzymes/non-enzymes classification model complexity based on composition, sequence, 3D and topological indices. *J. Theor. Biol.*, **254**(2):476-482. [doi:10.1016/j.jtbi.2008.06.003]

Nandy, A., 1994. A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to *globin* genes. *Curr. Sci.*, **66**(10):309-314.

Nandy, A., 1996. Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences. *Comput. Appl. Biosci.*, **12**(1):55-62.

Nandy, A., Basak, S.C., 2000. Simple numerical descriptor for quantifying effect of toxic substances on DNA sequences. *J. Chem. Inform. Comput. Sci.*, **40(4)**:915-919.

Nandy, A., Harle, M., Basak, S.C., 2006. Mathematical descriptors of DNA sequences: development and applications. *ARKIVOC*, **ix**:211-238.

Nei, M., Kumar, S., 2002. Molecular Evolution and Phylogenetics. Higher Education Press, Beijing, p.1-14 (in Chinese).

Pham, T.D., Zuegg, J., 2004. A probabilistic measure for alignment-free sequence comparison. *Bioinformatics*, **20**(18):3455-3461. [doi:10.1093/bioinformatics/bth426]

Randić, M., 2003. Condensed representation of DNA primary sequences. *J. Chem. Infrom. Comput. Sci.*, **40**(1):50-56.

Randić, M., 2007. 2-D Graphical representation of proteins based on physico-chemical properties of amino acids. *Chem. Phys. Lett.*, **440**(4-6):291-295. [doi:10.1016/j.cplett.2007.04.037]

Randić, M., Krilov, G., 1997. Characterization of 3-D sequences of proteins. *Chem. Phys. Lett.*, **272**(1-2):115-119. [doi:10.1016/S0009-2614(97)00447-8]

Randić, M., Balaban, A.T., 2003. On a four-dimensional representation of DNA primary sequences. *Chem. Inform. Comput. Sci.*, **43**(2):532-539.

Randić, M., Guo, X., Basak, S.C., 2001. On the characterization of DNA primary sequences by triplet of nucleic acid bases. *J. Chem. Inform. Comput. Sci.*, **41**(3):619-626.

Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison—a review. *Bioinformatics*, **19**(4):513-523. [doi:10.1093/bioinformatics/btg005]

Wen, J., Zhang, Y., 2009. A 2D graphical representation of protein sequence and its numerical characterization. *Chem. Phys. Lett.*, **476**(4-6):281-286. [doi:10.1016/j.cplett.2009.06.017]

Yao, Y., Dai, Q., Li, C., He, P., Nan, X., Zhang, Y., 2008. Analysis of similarity/dissimilarity of protein sequences. *Proteins*, **73**(4):864-871. [doi:10.1002/prot.22110]

Yao, Y., Dai, Q., Li, L., Nan, X., He, P., Zhang, Y., 2009. Similarity/dissimilarity studies of protein sequences based on a new 2D graphical representation. *J. Comput. Chem.*, **31**(5):1045-1052.

Yau, S.S.T., Yu, C., He, R., 2008. A protein map and its application. *DNA Cell Biol.*, **27**(5):241-250. [doi:10.1089/dna.2007.0676]

Zhang, C.T., Zhang, R., Ou, H.Y., 2003. The Z curve database: a graphic representation of genome sequences. *Bioinformatics*, **19**(5):593-599. [doi:10.1093/bioinformatics/btg041]