JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

A Novel Cylindrical Representation for Characterizing Intrinsic Properties of Protein Sequences

Jia-Feng Yu,*[,†,‡] Xiang-Hua Dou,[†] Hong-Bo Wang,[†] Xiao Sun,[‡] Hui-Ying Zhao,[§] and Ji-Hua Wang*[,†,∥]

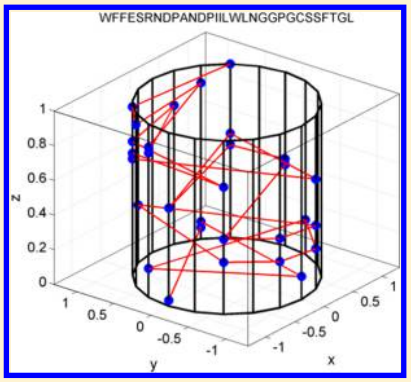
[†]Shandong Provincial Key Laboratory of Functional Macromolecular Biophysics, Institute of Biophysics, Dezhou University, Dezhou 253023, China

[‡]State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China

[§]Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland 4000, Australia

[∥]College of Physics and Electronic Information, Dezhou University, Dezhou 253023, China

ABSTRACT: The composition and sequence order of amino acid residues are the two most important characteristics to describe a protein sequence. Graphical representations facilitate visualization of biological sequences and produce biologically useful numerical descriptors. In this paper, we propose a novel cylindrical representation by placing the 20 amino acid residue types in a circle and sequence positions along the z axis. This representation allows visualization of the composition and sequence order of amino acids at the same time. Ten numerical descriptors and one weighted numerical descriptor have been developed to quantitatively describe intrinsic properties of protein sequences on the basis of the cylindrical model. Their applications to similarity/dissimilarity analysis of nine ND5 proteins indicated that these numerical descriptors are more effective than several classical numerical matrices. Thus, the cylindrical representation obtained here provides a new useful tool for visualizing and charactering protein sequences. An online server is available at http://biophy.dzu.edu.cn:8080/CNumD/input.jsp.



INTRODUCTION

The rapid development of high-throughput sequencing techniques has led to an explosive increase in biological sequences. How to process the vast amount of sequence data efficiently and effectively has been an urgent challenge. A DNA sequence consists of four letters (A, T, C, and G), whereas a protein sequence is made of 20 letters (A, R, N, D, C, Q, E, G, I, H, L, K, M, F, P, S, T, W, Y, and V). From such an abstract sequence of letters it is difficult to produce useful information directly. Graphical approaches are increasingly favored to provide intuitive and useful insights in understanding the mechanisms of biological processes[1−12] such as enzymatic reactions,[1−3] protein folding kinetics,[4] and drug metabolism.[5] One example is the use of the elegant wenxiang graphs[6] to analyze protein−protein interactions.[7] The first graphical representation was proposed for DNA sequences nearly three decades ago.[13] Since then, many different models have been developed.[14−23] By comparison, graphical representations of protein sequences have emerged only recently[24−39] because protein sequences have many more letters than DNA sequences. Jeffrey,[24] Basu et al.,[25] Yu et al.,[26] and Randić et al.[27] first reported the use of chaos game representation (CGR) for DNA and protein sequences. Randić outlined a two-dimensional (2D) graphical representation of proteins based on physicochemical properties of amino acids.[28] This method assigns 20 amino acids to 20 fixed vertices and presents a static graphical representation. Yao et al.[29] introduced a dynamic 2D graphical representation of protein sequences based on pK_a values that has an ability to recognize similarities among different proteins. Randić and co-workers further used spectral-like and zigzag representations[30] and starlike graphs[31] to encode DNA or protein sequences. Li et al.[32] proposed a 2D representation involving a 60D vector with five reduced amino acid types. Other representations for protein sequences (virtual genetic code and color map[33] and spherical representation[34]) have also been proposed. One advantage of graphical representation is the transformation of a linear sequence into a zigzag curve and numerical descriptors that makes the sequence more interpretable. Several versatile models and matrices such as Chou's PseAAC,[40,41] and the L/L and ${}^kL/{}^kL$ matrices[15] have been proposed to capture key protein features. Chou's PseAAC method, a mathematical model that numerically characterizes the protein sequence, was found to be useful in many aspects of computational proteomics[42−48] and has been extended to analysis of DNA sequences.[44−46] L/L and ${}^kL/{}^kL$ matrices[15] were proposed to transform the zigzag curves of biological sequences into numerical matrices, and their eigenvalues can be used as numerical descriptors for corresponding sequences. These matrices have been widely used in many graphical representation works.[36] Although these methods have been proved useful and are widely used, their calculation is complicated and time-consuming. Therefore, alternative numer-

Received: September 22, 2014
Published: May 6, 2015

DOI: 10.1021/ci500577m
J. Chem. Inf. Model. 2015, 55, 1261−1270

Table 1. Sequence Accession Numbers of the Nine ND5 Proteins in the Data Set

human	gorilla	common chimpanzee	pigmy chimpanzee	fin whale	blue whale	rat	mouse	opossum
YP_003024036	NP_008222	NP_008196	NP_008209	NP_006899	NP_007066	AP_004902	NP_904338	NP_007105

ical descriptors derived from the geometrical centers of the zigzag curves have been shown to be more convenient and efficient.[49]

Similarity between two protein sequences is an important parameter to reveal the relation between them. As a result, it is often employed to evaluate graphical representations. Most sequence alignment tools detect DNA/protein similarities on the basis of scoring matrices.[50,51] Ray developed a graphical method (MAVL/StickWRLD) that involves wrapping a positional weight matrix around a cylinder in order to detect interpositional dependences within DNA and protein sequence alignments.[52−54] Kultys et al.[55] proposed a technique to visualize and discover sequence motifs by aligning multiple sequences in physicochemical properties of amino acids. Sakai and Aerts[56] established a similar method based on sequence diversity diagrams for multiple sequence alignments.

Inspired by the methods developed above, in this work we have developed a novel cylindrical representation to characterize intrinsic features of protein sequences. This representation arranges the 20 amino acid types on the bottom circle of the cylinder. Each amino acid type forms a line on the cylindrical surface. Such a geometrical construction exhibits both the composition and distribution of amino acid residue types in a protein sequence. We have further proposed several numerical descriptors based on the cylindrical model for quantitative characterization of protein sequences. The application of the model to nine NADH dehydrogenase subunit 5 (ND5) proteins indicated that this novel graphical model is a convenient and efficient way to analyze protein sequences.

METHODS AND DATA SET

Data Set. We employed nine NADH dehydrogenase subunit 5 (ND5) proteins to demonstrate the present method because the data set has been widely utilized in many other graphical representation studies.[29,39] The accession numbers of these nine proteins are listed in Table 1.

Construction of the Cylindrical Representation. We employed cylindrical coordinates to display protein sequences on the surface of a unit cylinder. The transformation between cylindrical coordinates and Cartesian coordinates is shown in eq 1:

$$x_n = \cos\left(\frac{2\pi}{20} i_n\right) \qquad y_n = \sin\left(\frac{2\pi}{20} i_n\right) \qquad z_n = \frac{n}{N} \tag{1}$$

where N is the length of the protein, n = 1, 2, 3, ..., N is the positional index of an amino acid along the protein sequence, and i_n = 0, 1, 2, ..., 19 is the residue type of the nth amino acid residue. On the bottom circle, the 20 amino acid types are arranged counterclockwise according to the order of their hydropathy indices (Table 2); amino acids with same hydropathy index are sorted alphabetically. Thus, as indicated in Figure 1, the x and y coordinates of each amino acid type are fixed on the bottom circle of the cylinder. Table 2 lists the order of the 20 amino acid types and their respective x and y coordinates. In addition, z is used to demonstrate the sequence position. It could be represented in various forms such as $z = n$ or $z = 1 - 1/n$. Here the definition $z = n/N$ was used, so each sequence is compressed to the same range ($0 < z \leq 1$). By means of eq 1, a protein sequence can be represented by a set of connected points

Table 2. Order and x and y Coordinates of the 20 Types of Amino Acids

amino acid	order	hydropathy index	x	y
I	0	4.5	1	0
V	1	4.2	0.9511	0.309
L	2	3.8	0.809	0.5878
F	3	2.8	0.5878	0.809
C	4	2.5	0.309	0.9511
M	5	1.9	0	1
A	6	1.8	−0.309	0.9511
G	7	−0.4	−0.5878	0.809
T	8	−0.7	−0.809	0.5878
S	9	−0.8	−0.9511	0.309
W	10	−0.9	−1	0
Y	11	−1.3	−0.9511	−0.309
P	12	−1.6	−0.809	−0.5878
H	13	−3.2	−0.5878	−0.809
D	14	−3.5	−0.309	−0.9511
E	15	−3.5	0	−1
N	16	−3.5	0.309	−0.9511
Q	17	−3.5	0.5878	−0.809
K	18	−3.9	0.809	−0.5878
R	19	−4.5	0.9511	−0.309

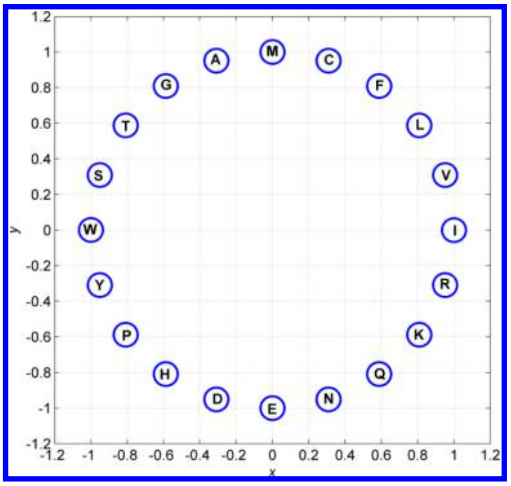


Figure 1. Arrangement of the 20 amino acid residue types on the unit circle.

$P = \{(x_1, y_1, z_1), (x_2, y_2, z_2), ..., (x_n, y_n, z_n)\}$. In the representation, there are 20 pillarlike columns on the cylindrical surface, each of which represents one amino acid residue type.

Cylindrical representations of the nine ND5 proteins are shown in Figure 2. The overall map of 20 amino acid columns and the lines connecting adjacent amino acids can be investigated. A close examination indicates that human, gorilla, common chimpanzee, and pigmy chimpanzee have similar contour maps. In addition, similarities between blue whale and fin whale and between rat and mouse can also be observed.

Figure 3 shows projections of the 3D cylindrical representations onto the 2D xy plane. The compositions of the amino acid residues in the nine ND5 proteins have a similar pattern according to the color-coded frequency shown in Figure 3.

DOI: 10.1021/ci500577m
J. Chem. Inf. Model. 2015, 55, 1261−1270

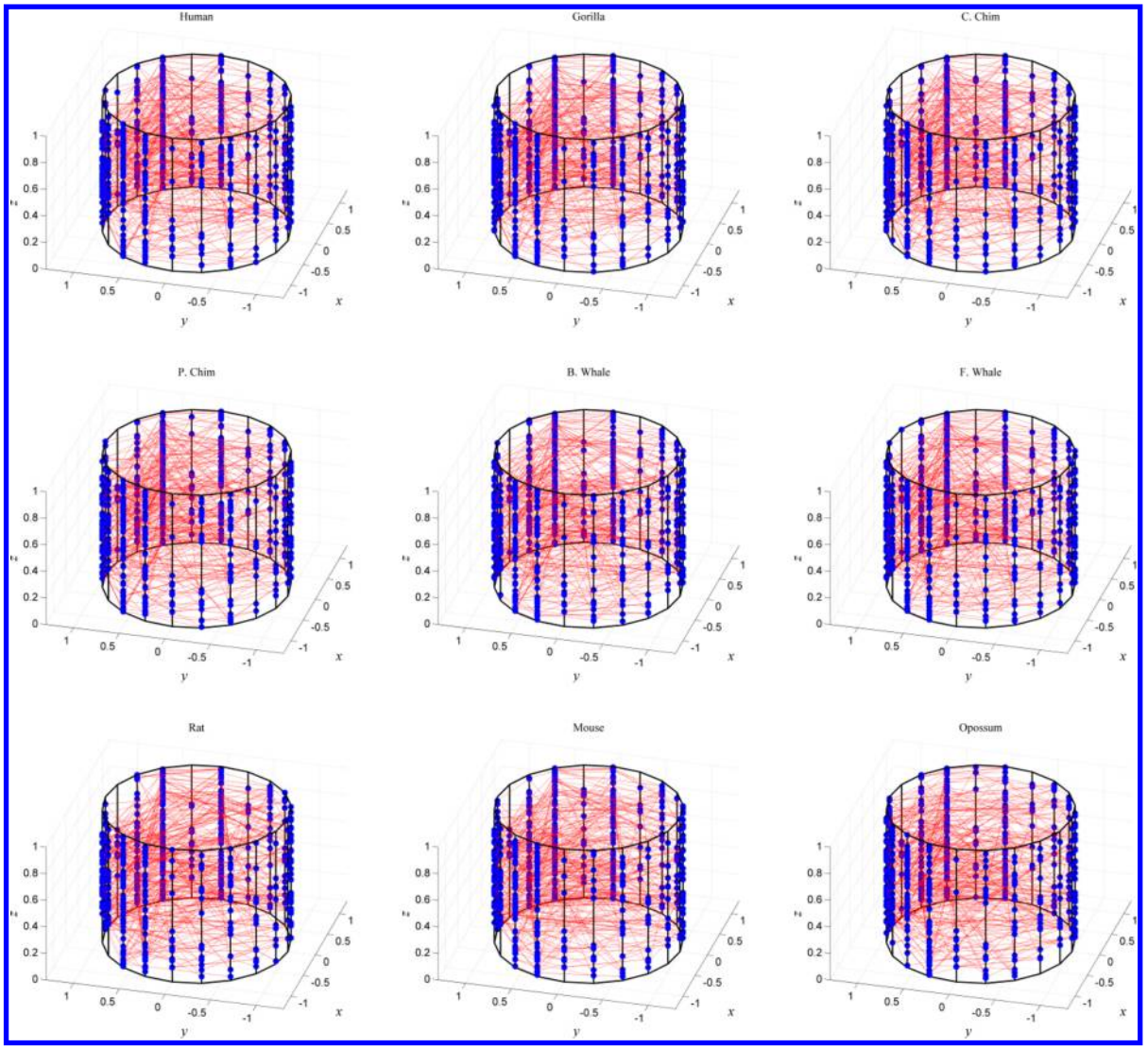


Figure 2. Cylindrical representations of the nine ND5 proteins.

Residue L has the highest frequency, while V, C, W, Y, H, D, E, Q, and R have the lowest usage. Some fine differences among different proteins can be detected from Figure 3. For example, the fractions of residue L in human, gorilla, common chimpanzee, and pigmy chimpanzee are 17.3%, 17.6%, 17.6%, and 17.6%, respectively, compared with 14.5% for opossum. Therefore, Figure 3 indicates that the cylindrical method provides a convenient tool for the visualization of amino acid compositions in protein sequences.

The sequence order of the amino acid residues is critical for a description of the intrinsic features of protein sequences. Integration of information on the sequence composition and order of amino acids has been successfully applied to many protein-related problems.[11,57] In Figure 4, we demonstrate that the cylindrical representation can further exhibit how amino residues are distributed along the primary sequence by decomposing each of the 3D representations in Figure 2 into four quadrants. Quadrant I has I, V, L, F, and C. As shown in quadrant I, human, gorilla, common chimpanzee, and pigmy chimpanzee share similar patterns. The same is true for blue whale and fin whale. Similar trends can be observed in quadrants II, III, and IV. Figure 4 presents a visualization tool showing how each amino acid type is utilized in the sequence by displaying each amino acid type along the cylindrical lines with dots and gaps.

Derivation of Numerical Descriptors Based on the Cylindrical Representation. Graphical representation methods allow the extraction of useful features. Here we define a covariance matrix $\mathbf{P}_1$ as

$$\mathbf{P}_1 = \begin{bmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{xy} & S_{yy} & S_{yz} \\ S_{xz} & S_{yz} & S_{zz} \end{bmatrix} \tag{2}$$

where

$$S_{xx} = \sum_n (x_n - \bar{x})^2 \qquad S_{xy} = \sum_n (x_n - \bar{x})(y_n - \bar{y})$$

$$S_{yy} = \sum_n (y_n - \bar{y})^2 \qquad S_{xz} = \sum_n (x_n - \bar{x})(z_n - \bar{z})$$

$$S_{zz} = \sum_n (z_n - \bar{z})^2 \qquad S_{yz} = \sum_n (y_n - \bar{y})(z_n - \bar{z})$$

in which $\bar{x}$, $\bar{y}$, and $\bar{z}$ are the geometrical centers, obtained as

DOI: 10.1021/ci500577m
J. Chem. Inf. Model. 2015, 55, 1261−1270

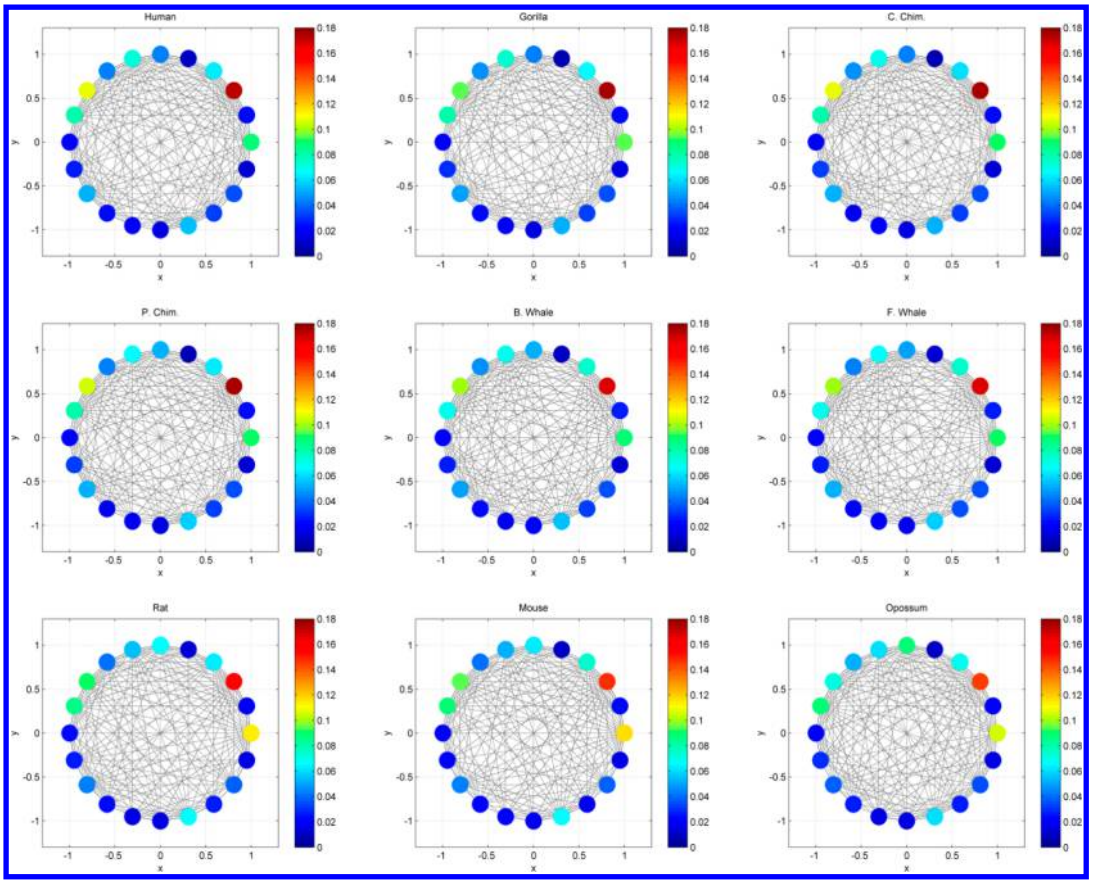


Figure 3. Projections of the cylindrical representations of the nine ND5 proteins onto the 2D *xy* plane.

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad \bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n \quad \bar{z} = \frac{1}{N}\sum_{n=1}^{N} z_n \tag{3}$$

We also define the vector $\mathbf{P_2} = [g_0, ..., g_i, ..., g_{19}]$ according to the geometrical centers of the 20 amino acid columns, given by

$$g_i = \frac{1}{N_i}\sum_{n_i=1}^{N_i} z_{n_i} \tag{4}$$

where $i = 0, 1, 2, ..., 19$ is the index for the 20 types of amino acids and N_i is the total number of amino acids in the ith column. To remove sequence-length dependence, two additional 20D vectors $\mathbf{P_3} = [g_0^{\mathrm{L}}, ..., g_i^{\mathrm{L}}, ..., g_{19}^{\mathrm{L}}]$ and $\mathbf{P_4} = [g_0^{\mathrm{N}}, ..., g_i^{\mathrm{N}}, ..., g_{19}^{\mathrm{N}}]$ are obtained by different normalization schemes as follows:

$$g_i^{\mathrm{L}} = \frac{g_i}{N} \tag{5}$$

$$g_i^{\mathrm{N}} = \frac{g_i}{\sum_i g_i} \tag{6}$$

In addition, we characterize the deviation of the amino acid residues from the geometrical center in each column by means of the vector $\mathbf{P_5} = [v_0, ..., v_i, ..., v_{19}]$, whose components are the standard deviations (SDs):

$$v_i = \sqrt{\frac{1}{N_i}\sum_{n_i}^{N_i} (z_{n_i} - g_i)^2}. \tag{7}$$

We can also employ the relative standard deviation (RSD) to obtain the vector $\mathbf{P_6} = [v_0^{\mathrm{R}}, ..., v_i^{\mathrm{R}}, ..., v_{19}^{\mathrm{R}}]$:

$$v_i^{\mathrm{R}} = \frac{1}{g_i}\sqrt{\frac{1}{N_i}\sum_{n_i}^{N_i} (z_{n_i} - g_i)^2} \tag{8}$$

The distribution disparities among the 20 amino acids can be described by the two matrices $\mathbf{P_7}$ and $\mathbf{P_8}$ based on eqs 4 and 7, respectively:

$$\mathbf{P_7} = \begin{bmatrix} g_{(0,0)} & \cdots & g_{(0,j)} & \cdots & g_{(0,19)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ g_{(i,0)} & \cdots & g_{(i,j)} & \cdots & g_{(i,20)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ g_{(19,0)} & \cdots & g_{(19,j)} & \cdots & g_{(19,19)} \end{bmatrix} \tag{9}$$

$$\mathbf{P_8} = \begin{bmatrix} v_{(0,0)} & \cdots & v_{(0,j)} & \cdots & v_{(0,19)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{(i,0)} & \cdots & v_{(i,j)} & \cdots & v_{(i,19)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{(19,0)} & \cdots & v_{(19,j)} & \cdots & v_{(19,19)} \end{bmatrix} \tag{10}$$

where $g_{(i,j)} = g_i - g_j$ and $v_{(i,j)} = v_i - v_j$ $(i, j = 0, 1, 2, ..., 19)$.

We also obtained the amino acid composition vector $\mathbf{P_9} = [f_0, ..., f_{19}]$ according the percentages of the 20 amino acids (f_i) as well as the vector $\mathbf{P_{10}} = [f_0, ..., f_{19}, g_0, ..., g_{19}]$ by combining $\mathbf{P_9}$ with $\mathbf{P_2}$.

The above descriptors except $\mathbf{P_1}$ depend on the z coordinate only. That is, they are independent of how the 20 amino acids are arranged on the circle. In addition, these descriptors are uniquely

DOI: 10.1021/ci500577m
J. Chem. Inf. Model. 2015, 55, 1261−1270

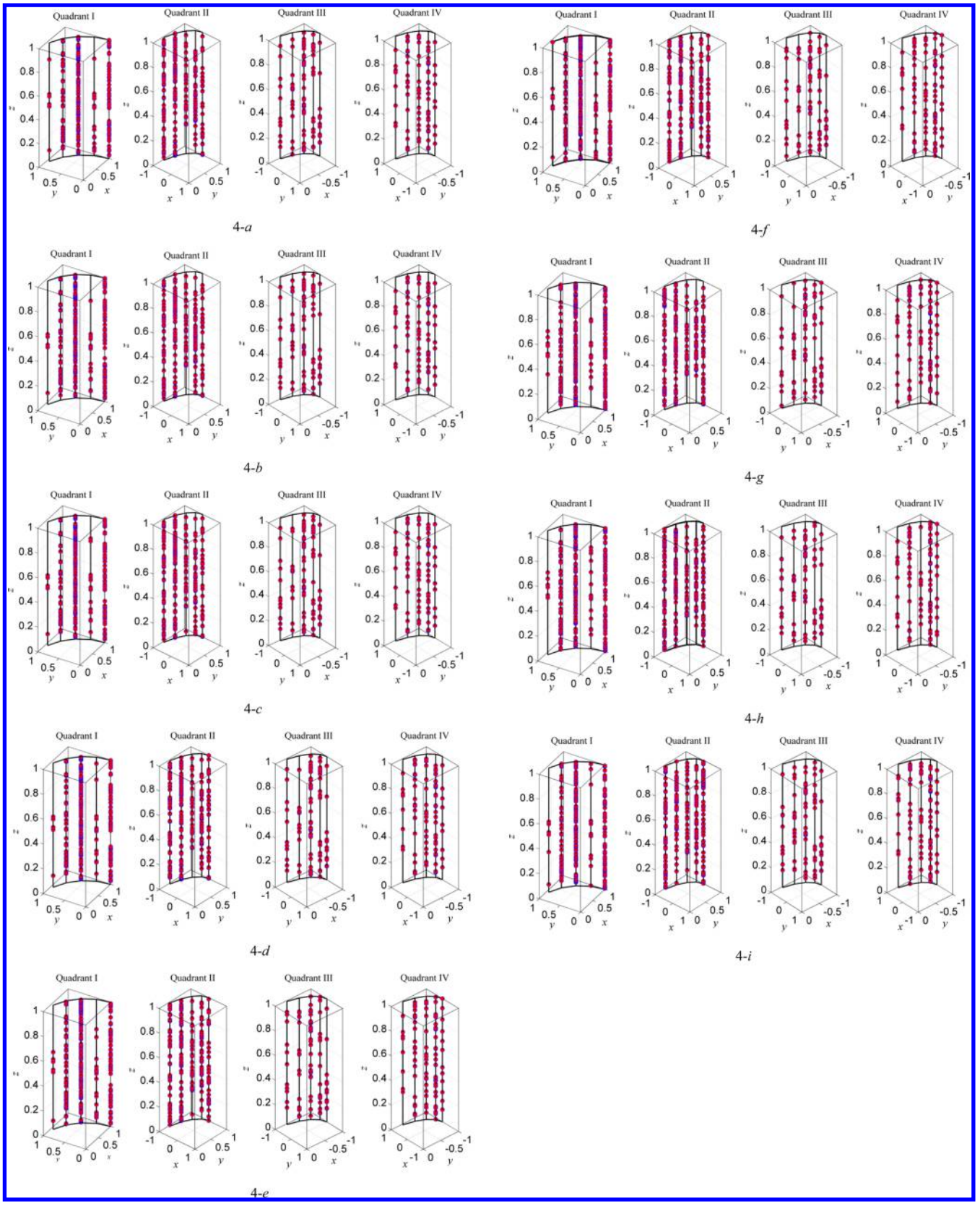


Figure 4. 3D representations of the nine ND5 proteins in four separate quadrants: (*a*) human; (*b*) gorilla; (*c*) common chimpanzee; (*d*) pigmy chimpanzee; (*e*) fin whale; (*f*) blue whale; (*g*) rat; (*h*) mouse; (*i*) opossum.

derived from the geometrical centers of the cylindrical representation and combine the compositions of the amino acid residues with their sequence position information. Therefore, they are simple and elegant.

DOI: 10.1021/ci500577m
J. Chem. Inf. Model. 2015, 55, 1261−1270

Measurement of Protein Sequence Similarity. Protein sequence similarity is used here to illlustrate the usefulness of the descriptors proposed above. The difference between two vectors can be evaluated in terms of the correlation angle and the Euclidean distance. The correlation angle θ between two sequences S_m and S_n is written as

$$\theta(S_m, S_n) = \arccos\frac{\sum_{t=1}^{20} v_t^{S_m} v_t^{S_n}}{\sqrt{\sum_{t=1}^{20} (v_t^{S_m})^2}\sqrt{\sum_{t=1}^{20} (v_t^{S_n})^2}} \quad (11)$$

where v^{S_m} and v^{S_n} are the numerical descriptors of S_m and S_n respectively. The Euclidean distance D is obtained as

$$D(S_m, S_n) = \sqrt{\sum_{t=1}^{20} (v_t^{S_m} - v_t^{S_n})^2} \quad (12)$$

A smaller value for either θ or D indicates a higher similarity between the two sequences.

APPLICATION CASES AND DISCUSSIONS

Mutation Analyses. We illustrate the usefulness of the cylindrical representation for mutation analysis using two short protein segments of the yeast *Saccharomyces cerevisiae*: "WTFESRNDPA" and "WFFESRNDPA". They have 10 residues with a mismatching amino acid at position 2. To facilitate visualization, the z coordinate is represented by $z = n$ rather than $z = n/N$. Figure 5 shows the cylindrical

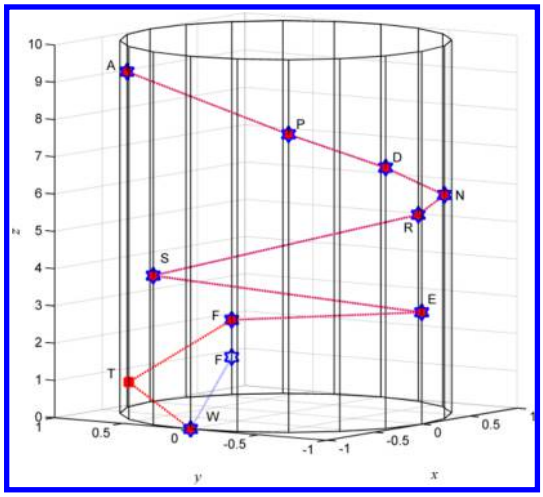


Figure 5. Cylindrical representations of two protein segments with a point mutation.

representations of the two segments. The mismatched amino acids in proteins I (red) and II (blue) are easy to detect. Figure 5 suggests that the cylindrical representation provides a convenient tool for visually inspecting similarity or dissimilarity among different sequences.

Similarity Analysis of Protein Sequences. Here we show that the descriptors $\mathbf{P}_1$–$\mathbf{P}_{10}$ can be used to measure the similarity among the nine ND5 proteins without pairwise alignment. We will demonstrate this by comparison with the results from the classical ClustalW2 program.[58] Our new descriptors are compared to previously employed descriptors: the versatile L/L and ${}^kL/{}^kL$ matrices. In this work, the top-10 leading eigenvalues of the L/L matrix and the top leading values of the ${}^kL/{}^kL$ matrix (k = 1, 2, 5, 10, 50) were used as numerical descriptors. Figure 6 compares the Pearson's correlation coefficients (PCCs) between

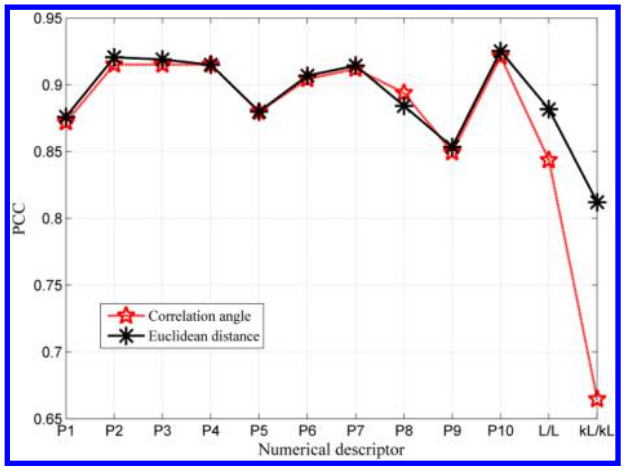


Figure 6. Pearson's correlation coefficients (PCC) between sequence similarities measured by ClustalW2 and those by various descriptors from this work ($\mathbf{P}_1$ to $\mathbf{P}_{10}$) and previous studies (L/L and ${}^kL/{}^kL$ matrices).

the results from ClustalW2 and those from various descriptors. The PCC is a measure of the correlation between two variables or matrices X and Y and gives a value between +1 and −1. The PCCs of $\mathbf{P}_1$–$\mathbf{P}_8$ are above 0.87 for both correlation angle and Euclidean distance. In particular, $\mathbf{P}_2$, $\mathbf{P}_3$, $\mathbf{P}_4$, $\mathbf{P}_6$, and $\mathbf{P}_7$ have PCC > 0.90 in terms of either the correlation angle or the Euclidean distance. The highest PCC is 0.92 for $\mathbf{P}_2$ based on either the correlation angle or the Euclidean distance. By comparison, the PCCs based on the L/L matrix are 0.84 for the correlation angle and 0.88 for the Euclidean distance. Even lower values are obtained for the ${}^kL/{}^kL$ matrix. P values of $\ll$0.01 for all of the numerical descriptors indicate results that are significantly consistent with those from ClustalW2. As an example, Tables 3 and 4 show the similarity/dissimilarity matrices of $\mathbf{P}_2$ for the nine ND5 protein sequences. As one can see from either table, there are higher similarities among the ND5 proteins of human, gorilla, common chimpanzee, and pigmy chimpanzee, between those of blue whale and fin whale, and between those of rat and mouse. The ND5 protein of opossum is not as similar to other the ND5 proteins. The results are consistent with evolutionary evidence.

Amino acid composition is one of the most efficient numerical parameters in many studies.[59] As shown in Figure 6, the PCCs of $\mathbf{P}_9$ are around 0.85 for both correlation angle and Euclidean distance, while those for $\mathbf{P}_2$ are 0.915 and 0.921, respectively. Interestingly, $\mathbf{P}_{10}$, the combination of $\mathbf{P}_2$ and $\mathbf{P}_9$, further improves the PCCs for both measurements to 0.921 and 0.925, respectively. This suggests that $\mathbf{P}_2$ and $\mathbf{P}_9$ contain complementary information. It is possible that a different weight might further improve the correlation. We define the descriptor $\mathbf{P}_{11}$ by

$$\mathbf{P}_{11} = \begin{cases} \alpha\mathbf{P}_9 & 0 < \alpha \le 1 \\ \beta\mathbf{P}_2 & 0 < \beta \le 1 \end{cases} \quad (13)$$

where α and β are the weighting coefficients. Figure 7 shows the PCCs between the results from ClustalW2 and $\mathbf{P}_{11}$ with different weighting coefficients varied with a step size of 0.1. The optimal values are 0.6 for α and 0.2 for β, with PCCs values of 0.941 for the correlation angle and 0.941 for the Euclidean distance. This illustrates the usefulness of the combined descriptors.

Biological Significance of the Cylindrical Representation. The above analyses illustrate that the proposed numerical descriptors based on our cylindrical representation are bio-

DOI: 10.1021/ci500577m
J. Chem. Inf. Model. 2015, 55, 1261−1270

Table 3. Similarity/Dissimilarity Matrix of P_2 Based on the Correlation Angle

	gorilla	c. chimp.	p. chimp.	f. whale	b. whale	rat	mouse	opossum
human	0.0607	0.0613	0.0729	0.0809	0.0829	0.117	0.1036	0.1289
gorilla		0.0459	0.0697	0.0833	0.0975	0.1317	0.1136	0.1125
c. chimp.			0.0373	0.0946	0.1031	0.1169	0.1015	0.0943
p. chimp.				0.1046	0.1174	0.117	0.1003	0.095
b. whale					0.0485	0.1125	0.0993	0.1308
f. whale						0.112	0.0986	0.1341
rat							0.0756	0.1456
mouse								0.1299

Table 4. Similarity/Dissimilarity Matrix of P_2 Based on the Euclidean Distance

	gorilla	c. chimp.	p. chimp.	f. whale	b. whale	rat	mouse	opossum
human	0.1326	0.134	0.1595	0.1781	0.1829	0.2643	0.2338	0.2922
gorilla		0.1003	0.1526	0.1834	0.2151	0.2967	0.2561	0.2576
c. chimp.			0.0823	0.2083	0.2276	0.2652	0.2308	0.22
p. chimp.				0.2299	0.2584	0.2638	0.2263	0.2186
b. whale					0.1069	0.2528	0.2227	0.2945
f. whale						0.2513	0.2221	0.3015
rat							0.1747	0.3267
mouse								0.2914

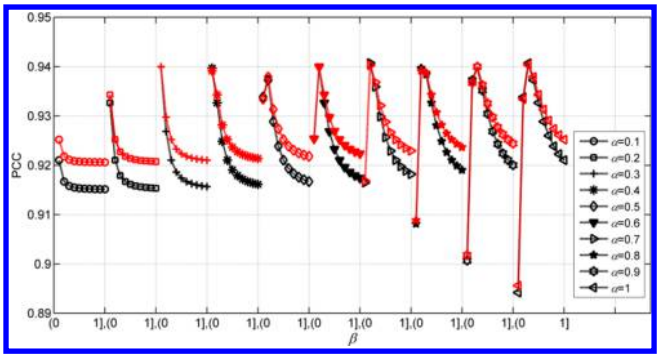


Figure 7. PCCs of P_{11} with different values of the weighting parameters. The black and red curves represent the results using the correlation angle and Euclidean distance, respectively.

logically meaningful. According to eqs 4 and 7, P_2 indicates the distribution center of each amino acid type and P_5 measures the deviation of each amino acid type from its distribution center in the protein sequence. Figure 8 shows the compositions of the amino acids in the nine ND5 proteins. There is no significance difference between the compositions of nine ND5 proteins. By comparison, the geometrical centers (P_2) and the standard deviations (P_5) (Figure 9) reveal the characteristic differences between different proteins signaled by W in the P_2 curves and M and Y in the P_5 curves. Some fine differences can also be observed among homologous species, such as residue M in the P_2 curves of common chimpanzee and pigmy chimpanzee and residue P in the P_5 curves of fin whale and blue whale. Thus, P_2 and P_5 provide effective and complementary features to discriminate highly homologous protein sequences.

CONCLUSION

Sequence analysis is one of the most important problems in bioinformatics. Graphical representations provide simple and convenient tools for qualitatively and quantitatively describing

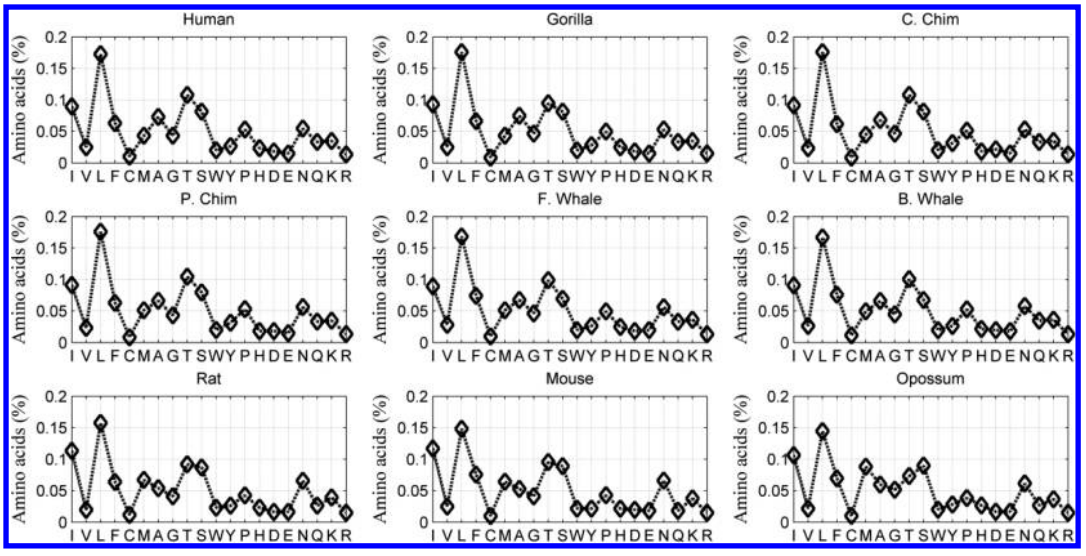


Figure 8. Compositions of the amino acid residues of the nine ND5 proteins.

DOI: 10.1021/ci500577m
J. Chem. Inf. Model. 2015, 55, 1261−1270

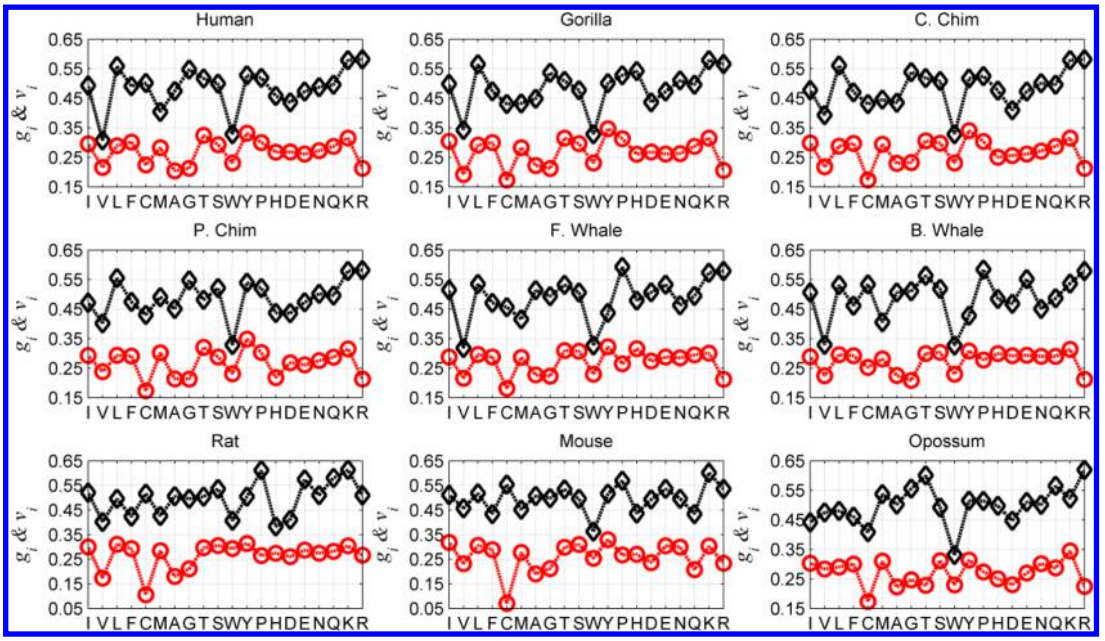


Figure 9. Geometrical centers ($\mathbf{P}_2$, black) and standard deviations ($\mathbf{P}_5$, red) of the 20 amino acid residues of the nine ND5 proteins.

intrinsic features of biological sequences.[60−64] In this paper, we have proposed a novel cylindrical representation by arranging the 20 amino acid types on a cylindrical surface according to their respective sequence positions. Its application to nine ND5 proteins indicates that this graphical model is effective for describing intrinsic properties of protein sequences. Numerical descriptors derived from the cylindrical representation can quantitatively describe the intrinsic characteristics of the composition and sequence order of the amino acid residues. These descriptors have been shown to be effective for detecting similarity and dissimilarity between two proteins without sequence alignment. Thus, the cylindrical representation is a useful new tool for protein sequence analysis. An online server has been set up to calculate the values for all of the numerical descriptors at http://biophy.dzu.edu.cn:8080/CNumD/input.jsp.

AUTHOR INFORMATION

Corresponding Authors

*Phone: +86 534 8982557. Fax: +86 534 8985884. E-mail: jfyu1979@126.com (J.-F.Y.).
*E-mail: jhw25336@126.com (J.-H.W.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Prof. Yaoqi Zhou (Griffith University) and the anonymous referees for their valuable suggestions that have improved this article. This work was supported by the National Natural Science Foundation of China (61302186 and 61271378), the Shandong Natural Science Foundation (ZR2010CQ041) and funding from the State Key Laboratory of Bioelectronics of Southeast University.

REFERENCES

(1) Chou, K. C.; Forsen, S. Graphical Rules for Enzyme-Catalyzed Rate Laws. *Biochem. J.* **1980**, *187*, 829−835.
(2) Zhou, G. P.; Deng, M. H. An Extension of Chou's Graphic Rules for Deriving Enzyme Kinetic Equations to Systems Involving Parallel Reaction Pathways. *Biochem. J.* **1984**, *222*, 169−176.
(3) Chou, K. C. Graphic Rules in Steady and Non-Steady Enzyme Kinetics. *J. Biol. Chem.* **1989**, *264*, 12074−12079.
(4) Chou, K. C. Review: Applications of Graph Theory to Enzyme Kinetics and Protein Folding Kinetics. Steady and Non-Steady State Systems. *Biophys. Chem.* **1990**, *35*, 1−24.
(5) Chou, K. C. Graphic Rule for Drug Metabolism Systems. *Curr. Drug Metab.* **2010**, *11*, 369−378.
(6) Chou, K. C.; Lin, W. Z.; Xiao, X. Wenxiang: A Web-Server for Drawing Wenxiang Diagrams. *Nat. Sci.* **2011**, *3*, 862−865.
(7) Zhou, G. P. The Disposition of the LZCC Protein Residues in Wenxiang Diagram Provides New Insights into the Protein−Protein Interaction Mechanism. *J. Theor. Biol.* **2011**, *284*, 142−148.
(8) Xiao, X.; Shao, S. H. A Probability Cellular Automaton Model for Hepatitis B Viral Infections. *Biochem. Biophys. Res. Commun.* **2006**, *342*, 605−610.
(9) Xiao, X.; Shao, S.; Ding, Y.; Huang, Z. An Application of Gene Comparative Image for Predicting the Effect on Replication Ratio by HBV Virus Gene Missense Mutation. *J. Theor. Biol.* **2005**, *235*, 555−565.
(10) Xiao, X.; Shao, S.; Ding, Y.; Huang, Z. Using Cellular Automata To Generate Image Representation for Biological Sequences. *Amino Acids* **2005**, *28*, 29−35.
(11) Xiao, X.; Wang, P. Predicting Protein Structural Classes with Pseudo Amino Acid Composition: An Approach Using Geometric Moments of Cellular Automaton Image. *J. Theor. Biol.* **2008**, *254*, 691−696.
(12) Xiao, X.; Wang, P. GPCR-CA: A Cellular Automaton Image Approach for Predicting G-Protein-Coupled Receptor Functional Classes. *J. Comput. Chem.* **2009**, *30*, 1414−1423.
(13) Hamori, E.; Ruskin, J. H Curves, a Novel Method of Representation of Nucleotide Series Especially Suited for Long DNA Sequences. *J. Biol. Chem.* **1983**, *258*, 1318−1327.
(14) Zhang, C. T.; Zhang, R. Analysis of Distribution of Bases in the Coding Sequences by a Diagrammatic Technique. *Nucleic Acids Res.* **1991**, *19*, 6313−6317.
(15) Randić, M.; Vračko, M.; Lerš, N.; Plavšić, D. Novel 2-D Graphical Representation of DNA Sequences and Their Numerical Characterization. *Chem. Phys. Lett.* **2003**, *368*, 1−6.
(16) Yau, S. S.; Wang, J.; Nikenjad, A.; Lu, C.; Jin, N.; Ho, Y. K. DNA Sequence Representation without Degeneracy. *Nucleic Acids Res.* **2003**, *31*, 3078−3080.

DOI: 10.1021/ci500577m
J. Chem. Inf. Model. 2015, 55, 1261−1270

(17) Yu, J. F.; Sun, X.; Wang, J. H. TN Curve: A Novel 3D Graphical Representation of DNA Sequence Based on Trinucleotides and Its Applications. *J. Theor. Biol.* **2009**, *261*, 459−468.
(18) Zhang, Z. J. DV-Curve: A Novel Intuitive Tool for Visualizing and Analyzing DNA Sequences. *Bioinformatics* **2009**, *25*, 1112−1117.
(19) Yu, J. F.; Wang, J. H.; Sun, X. Analysis of Similarities/Dissimilarities of DNA Sequences Based on a Novel Graphical Representation. *MATCH: Commun. Math. Comput. Chem.* **2010**, *63*, 493−512.
(20) Aram, V.; Iranmanesh, A. 3D-Dynamic Representation of DNA Sequences. *MATCH: Commun. Math. Comput. Chem.* **2012**, *67*, 809−816.
(21) Wąż, P.; Bielińska-Wąż, D. 3D-Dynamic Representation of DNA Sequences. *J. Mol. Model.* **2014**, *20*, 2141.
(22) Jeong, B. S.; Baria, A. T. M. G.; Reaz, M. R.; Jeon, S.; Lima, C. G.; Choi, H. J. Codon-Based Encoding for DNA Sequence Analysis. *Methods* **2014**, *67*, 373−379.
(23) Jafarzadeh, N.; Iranmanesh, A. C-Curve: A Novel 3D Graphical Representation of DNA Sequence Based on Codons. *Math. Biosci.* **2013**, *241*, 217−224.
(24) Jeffrey, H. J. Chaos Game Representation of Gene Structure. *Nucleic Acids Res.* **1990**, *18*, 2163−2170.
(25) Basu, S.; Pan, A.; Dutta, C. Chaos Game Representation of Proteins. *J. Mol. Graphics Modell.* **1997**, *15*, 279−289.
(26) Yu, Z. G.; Anh, V.; Lau, K. S. Chaos Game Representation of Protein Sequences Based on the Detailed HP Model and Their Multifractal and Correlation Analyses. *J. Theor. Biol.* **2004**, *226*, 341−348.
(27) Randić, M.; Butina, D.; Zupan, J. Novel 2-D Graphical Representation of Proteins. *Chem. Phys. Lett.* **2006**, *419*, 528−532.
(28) Randić, M. 2-D Graphical Representation of Proteins Based on Physico-Chemical Properties of Amino Acids. *Chem. Phys. Lett.* **2007**, *444*, 176−180.
(29) Yao, Y. H.; Dai, Q.; Li, C.; Nan, X. Y.; Zhang, Y. Z. Analysis of Similarity/Dissimilarity of Protein Sequences. *Proteins* **2008**, *73*, 864−871.
(30) Zupan, J.; Randić, M. Algorithm for Coding DNA Sequences into "Spectrum-like" and "Zigzag" Representations. *J. Chem. Inf. Model.* **2005**, *45*, 309−313.
(31) Randić, M.; Zupan, J.; Vikić-Topić, D. On Representation of Proteins by Star-like Graphs. *J. Mol. Graphics Modell.* **2007**, *26*, 290−305.
(32) Li, C.; Yu, X. Q.; Yang, L.; Zheng, X.; Wang, Z. 3-D Maps and Coupling Numbers for Protein Sequences. *Physica A* **2009**, *388*, 1967−1972.
(33) Randić, M.; Mehulić, K.; Vukicević, D.; Pisanski, T.; Vikić-Topić, D.; Plavsić, D. Graphical Representation of Proteins as Four-Color Maps and Their Numerical Characterization. *J. Mol. Graphics Modell.* **2009**, *27*, 637−641.
(34) Abo el Maaty, M. I.; Abo-Elkhier, M. M.; Abd Elwahaab, M. A. Representation of Protein Sequences on Latitude-like Circles and Longitude-like Semi-circles. *Chem. Phys. Lett.* **2010**, *493*, 386−391.
(35) Wu, Z. C.; Xiao, X.; Chou, K. C. 2D-MH: A Web-Server for Generating Graphic Representation of Protein Sequences Based on the Physicochemical Properties of Their Constituent Amino Acids. *J. Theor. Biol.* **2010**, *267*, 29−34.
(36) Yu, J. F.; Sun, X.; Wang, J. H. A Novel 2D Graphical Representation of Protein Sequence Based on Individual Amino Acid. *Int. J. Quantum Chem.* **2011**, *111*, 2835−2843.
(37) Li, Z.; Geng, C.; He, P. A.; Yao, Y. H. A Novel Method of 3D Graphical Representation and Similarity Analysis for Proteins. *MATCH: Commun. Math. Comput. Chem.* **2014**, *71*, 213−226.
(38) Li, Y. H.; Liu, Q.; Zheng, X. Q.; He, P. A. UC-Curve: A Highly Compact 2D Graphical Representation of Protein Sequences. *Int. J. Quantum Chem.* **2014**, *114*, 409−415.
(39) Yao, Y. H.; Yan, S. J.; Han, J. N.; Dai, Q.; He, P. A. A Novel Descriptor of Protein Sequences and Its Application. *J. Theor. Biol.* **2014**, *347*, 109−117.
(40) Chou, K. C. Prediction of Protein Cellular Attributes Using Pseudo Amino Acid Composition. *Proteins* **2001**, *43*, 246−255.
(41) Chou, K. C. Using Amphiphilic Pseudo Amino Acid Composition To Predict Enzyme Subfamily Classes. *Bioinformatics* **2005**, *21*, 10−19.
(42) Chou, K. C. Some Remarks on Protein Attribute Prediction and Pseudo Amino Acid Composition. *J. Theor. Biol.* **2011**, *273*, 236−247.
(43) Du, P.; Wang, X.; Xu, C.; Gao, Y. PseAAC-Builder: A Cross-Platform Stand-Alone Program for Generating Various Special Chou's Pseudo-Amino Acid Compositions. *Anal. Biochem.* **2012**, *425*, 117−119.
(44) Chen, W.; Feng, P.-M.; Lin, H.; Chou, K.-C. iRSpot-PseDNC: Identify Recombination Spots with Pseudo Dinucleotide Composition. *Nucleic Acids Res.* **2013**, *41*, No. e68.
(45) Qiu, W. R.; Xiao, X. iRSpot-TNCPseAAC: Identify Recombination Spots with Trinucleotide Composition and Pseudo Amino Acid Components. *Int. J. Mol. Sci.* **2014**, *15*, 1746−1766.
(46) Lin, H.; Deng, E. Z.; Ding, H.; Chen, W. iPro54-PseKNC: A Sequence-Based Predictor for Identifying Sigma-54 Promoters in Prokaryote with Pseudo *k*-Tuple Nucleotide Composition. *Nucleic Acids Res.* **2014**, *42*, 12961−12972.
(47) Chen, W.; Zhang, X.; Brooker, J.; Lin, H.; Zhang, L. Q.; Chou, K. C. PseKNC-General: A Cross-Platform Package for Generating Various Modes of Pseudo Nucleotide Compositions. *Bioinformatics* **2014**, *31*, 119−120.
(48) Chen, W.; Lei, T. Y.; Jin, D. C.; Lin, H.; Chou, K. C. PseKNC: A Flexible Web-Server for Generating Pseudo *k*-Tuple Nucleotide Composition. *Anal. Biochem.* **2014**, *456*, 53−60.
(49) Liu, Z. B.; Liao, B.; Zhu, W.; Huang, G. H. A 2D Graphical Representation of DNA Sequence Based on Dual Nucleotides and Its Application. *Int. J. Quantum Chem.* **2009**, *109*, 948−958.
(50) Randić, M. Very Efficient Search for Protein Alignment—VESPA. *J. Comput. Chem.* **2014**, *33*, 702−707.
(51) Randić, M.; Pisanski, T. Protein Alignment: Exact versus Approximate. An Illustration. *J. Comput. Chem.* **2015**, *36*, 1069−1074.
(52) Ray, R. C. MAVL and Stickwrld: Visually Exploring Relationships in Nucleic Acid Sequence Alignments. *Nucleic Acids Res.* **2004**, *32*, W59−W63.
(53) Ray, R. C. MAVL/Stickwrld for Protein: Visualizing Protein Sequence Families To Detect Non-consensus Features. *Nucleic Acids Res.* **2005**, *33*, W315−W319.
(54) Ozer, H. G.; Ray, R. C. MAVL/Stickwrld: Analyzing Structural Constraints Using Interpositional Dependencies in Biomolecular Sequence Alignments. *Nucleic Acids Res.* **2005**, *34*, W133−W136.
(55) Kultys, M.; Nicholas, L.; Schwarz, R.; Goldman, N.; King, J. Sequence Bundles: A Novel Method for Visualising, Discovering and Exploring Sequence Motifs. *BMC Proc.* **2014**, *8* (Suppl. 2), No. S8.
(56) Sakai, R.; Aerts, J. Sequence Diversity Diagram for Comparative Analysis of Multiple Sequence Alignments. *BMC Proc.* **2014**, *8* (Suppl. 2), No. S9.
(57) Chou, K. C.; Shen, H. B. Recent Progress in Protein Subcellular Location Prediction. *Anal. Biochem.* **2007**, *370*, 1−16.
(58) Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J.; Higgins, D. G. Clustal W and Clustal X version 2.0. *Bioinformatics* **2007**, *23*, 2947−2948.
(59) Weathers, E. A.; Paulaitis, M. E.; Woolf, T. B.; Hoh, J. H. Reduced Amino Acid Alphabet Is Sufficient To Accurately Recognize Intrinsically Disordered Protein. *FEBS Lett.* **2004**, *576*, 348−352.
(60) Zhang, C. T.; Zhang, R. An Isochore Map of the Human Genome Based on the Z Curve Method. *Gene* **2003**, *317*, 127−135.
(61) Guo, F. B.; Ou, H. Y.; Zhang, C. T. ZCURVE: A New System for Recognizing Protein-Coding Genes in Bacterial and Archaeal Genomes. *Nucleic Acids Res.* **2003**, *31*, 1780−1789.
(62) Chen, L. L.; Ma, B. G.; Gao, N. Reannotation of Hypothetical ORFs in Plant Pathogen *Erwinia carotovora* subsp. *atroseptica* SCRI1043. *FEBS J.* **2008**, *275*, 198−206.
(63) Yu, J. F.; Sun, X. Reannotation of Protein-Coding Genes Based on an Improved Graphical Representation of DNA Sequence. *J. Comput. Chem.* **2010**, *31*, 2126−2135.

DOI: 10.1021/ci500577m
J. Chem. Inf. Model. 2015, 55, 1261−1270

(64) Liao, B.; Liao, B. Y.; Sun, X. M.; Zeng, Q. G. A Novel Method for Similarity Analysis and Protein Sub-cellular Localization Prediction. *Bioinformatics* **2010**, *26*, 2678−2683.

DOI: 10.1021/ci500577m
J. Chem. Inf. Model. 2015, 55, 1261−1270