

Research Article

Analysis of Similarity/Dissimilarity of DNA Sequences Based on Chaos Game Representation

Wei Deng^{1,2} and Yihui Luan¹

¹ School of Mathematics, Shandong University, Jinan 250100, China

² School of Science, Shandong Jianzhu University, Jinan 250101, China

Correspondence should be addressed to Yihui Luan; yhluan@sdu.edu.cn

Received 20 November 2012; Revised 27 January 2013; Accepted 8 February 2013

Academic Editor: Yong Zhang

Copyright © 2013 W. Deng and Y. Luan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The Chaos Game is an algorithm that can allow one to produce pictures of fractal structures. Considering that the four bases A, G, C, and T of DNA sequences can be divided into three classes according to their chemical structure, we propose different kinds of CGR-walk sequences. Based on CGR coordinates of random sequences, we introduce some invariants for the DNA primary sequences. As an application, we can make the examination of similarity/dissimilarity among the first exon of β -globin gene of different species. The results indicate that our method is efficient and can get more biological information.

1. Introduction

A DNA sequence is comprised of four different nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). Since the DNA molecule contains plentiful biological, physical, and chemical information, it has become very important to analyze DNA sequences statistically. Now the nucleotides stored in GenBank have exceeded hundreds of millions of bases and the increasing rate is considerably rapid. Therefore, biologists, physicists, mathematicians, and computer specialists have adopted different techniques to research DNA sequences in recent years, including the statistical methods and some mapping rules of the bases.

A great number of studies have proposed to convert the DNA sequences into digital sequences before downstream analysis. There are many statistical methods such as random walk, lévy-walk, entropy near method, root-mean-square fluctuation, wavelet transform and Fourier transform, and so forth, [1–12], which can be used as effective tools to process the DNA sequences. One-dimensional DNA walk was first proposed by Peng et al. [1]. Bai et al. [13] later discussed the representation of DNA primary sequences by the same walk. Meanwhile, some investigators proposed several kinds of graphical representation of DNA sequences from different perspectives. For example, G-curve and H-curve were first

proposed by Hamori and Ruskin in 1983 [14]. R. Zhang and C. T. Zhang [15] considered a DNA primary sequence termed as Z-curve. Several researchers in their recent studies have outlined different kinds of graphical representation of DNA sequences based on 2D [16–21], 3D [22–25], 4D [26], 5D [27], and 6D [28] spaces. We here need to stress Chaos Game Representation (CGR) which was proposed as a scale-independent representation for genomic sequences by Jeffrey [3] in 1990. Gao and Xu [29] pointed out that the CGR-walk model can easily generate a model sequence and can be fitted with a long-memory ARFIMA (p, d, q) model reasonably. However, they treated the four bases equally and ignored the hidden chemical classification of nucleotides.

Motivated by the above work, we consider in this paper different classifications of the four bases according to their chemical structure and the strength of the hydrogen bond, that is, purine $R = \{A, G\}$ and pyrimidine $Y = \{C, T\}$; amino group $M = \{A, C\}$ and keto group $K = \{G, T\}$; weak H-bonds $W = \{A, T\}$ and strong H-bonds $S = \{G, C\}$. Then we give three kinds of mapping from the four bases A, C, G, and T to the continuous space and reconstruct CGR-walk sequences based on CGR coordinates. So we can convert a DNA sequence into a random numeric sequence, then select some numerical characterizations of the random sequence as new invariants for the DNA sequence. As an application,

we make a comparison of the similarity and dissimilarity of the first exon of β -globin gene sequences derived from nine species.

2. CGR-Walk Based on Three kinds of Classification and Primary Sequences

2.1. The CGR Space Proposed by Jeffrey. During the past several years, a new field of physics has developed, known as “nonlinear dynamics,” “chaotic dynamical systems,” or simply “chaos.” In fact, the technique of CGR, formally an iterative mapping, can be traced further back to the foundation of statistical mechanics, in particular, to chaos theory [2]. Based on the technique from chaotic dynamics, CGR produces a picture of gene sequence which displays both local and global patterns. The Chaos Game is an algorithm which allows one to produce pictures of fractal structures. Mathematically, it is described by an iterated function system (IFS).

The CGR space can be viewed as a continuous reference system, where all possible sequences of any length occupy a unique position. And the position is produced by the four possible nucleotides, which are treated as vertices of a binary square. So it is planar. Since a genetic sequence can be treated formally as a string composed of the four letters “A,” “C,” “G,” and “T” (or “U”), the binary CGR vertices are assigned to the four nucleotides as $A = (0, 0)$, $G = (1, 1)$, $C = (0, 1)$, $T = (1, 0)$. The CGR coordinates are calculated iteratively by moving a pointer to half the distance between the previous position and the current binary representation. For example, if a “G,” is the next base, then a point is plotted half way between the previous point and the “G” corner. The iterated function can be given by

$$CGR_i = CGR_{i-1} - 0.5 (CGR_{i-1} - g_i), \quad (1)$$

where

$$i = 1, \dots, n_G; \quad CGR_0 = (0.5, 0.5); \quad g_i \in \{A, G, C, T\}. \quad (2)$$

We take the first 6 bases of the sequence of human β -globin in Table 1 as an example and present the above procedure in Figure 1.

2.2. The Newly Proposed CGR Space. The aforementioned work treats the four nucleic acid bases equally. In this paper, however, we take the chemical structures of the four nucleic acid bases into consideration and make adjustments to the classification based on the elements of the minor diagonal. In the CGR space proposed by Jeffrey, the elements of the minor diagonal are purine $R = \{A, G\}$ and the leading diagonal elements are pyrimidine $Y = \{C, T\}$. Considering amino group $M = \{A, C\}$ and keto group $K = \{G, T\}$, we get the second CGR space as shown in Figure 2. In the same way, according to the strength of the hydrogen bond, the bases can also be classified into weak H-bonds $W = \{A, T\}$ and strong H-bonds $S = \{G, C\}$, so the third kind of CGR space is obtained in Figure 3.

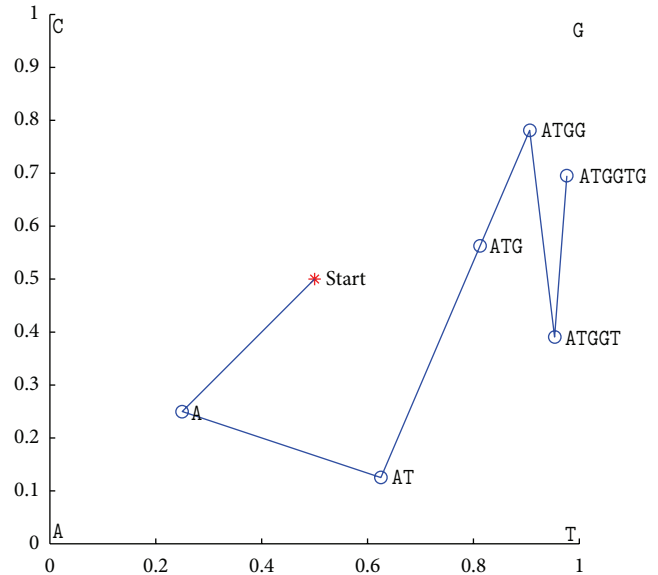


FIGURE 1: CGR-RY of the first 6 bases of exon-1 of human β -globin: ATGGTG.

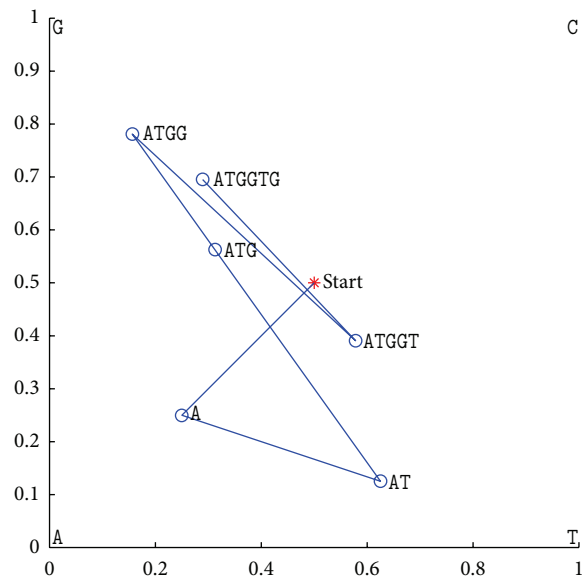


FIGURE 2: CGR-MK of the first 6 bases of exon-1 of human β -globin: ATGGTG.

2.3. CGR-Walk Digital Sequence. Now we can obtain map relationships between DNA sequences and the CGR coordinates in a right-angled plane. For a DNA sequence, we define an equation as follows:

$$z_i = x_i + y_i, \quad (3)$$

where x_i and y_i are the x -coordinate and y -coordinate of CGR, respectively. Then we can get a data sequence $\{z_i : i = 1, 2, \dots, N\}$. In this way, we convert a DNA sequence into a random walk sequence under three different patterns. Consistent with the above three figures, we call them CGR-RY-, CGR-MK-, and CGR-WS-walk sequences, respectively.

TABLE 1: The coding sequences of the first exon of β -globin gene of different species.

Species	Coding sequence
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGT TACTGCCCTGTGGGGCAAGGTGAACGTGGATTAAG TTGGTGGTGAGGCCCTGGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGG CTTCTGGGGCAAGGTGAAAGTGGATGAAGTTGGTG CTGAGGCCCTGGGCAG
Opossum	ATGGTGCACCTTGACTTCTGAGGAGAAGAAGTGC TCACTACCATCTGGTCTAAGGTGCAGGTTGACCA GACTGGTGGTGAGGCCCTGGGCAG
Gallus	ATGGTGCACCTGGACTGCTGAGGAGAAGCAGCTCAT CACCGCCCTCTGGGGGAAGGTCAATGTGGCCGAAT GTGGGGCCGAAGCCCTGGCCAG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGT CACCTCTCTGTGGGGCAAGGTGGATGTAGAGAAAAG TTGGTGGCGAGGCCCTGGGCAG
Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTG TCTCTTGCCCTGTGGGCAAAGGTGAACCCCGATGAA GTTGGTGGTGAGGCCCTGGGCAGG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGT CACTGCCCTGTGGGGCAAGGTGAATGTGGAAGAAG TTGGTGGTGAGGCCCTGGGC
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGT TAGTGGCCTGTGGGGAAAGGTGAACCCTGATAATG TTGGCGCTGAGGCCCTGGGCAG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGT TACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG TTGGTGGTGAGGCCCTGGGCAGG

TABLE 2: Hurst exponent of the CGR-walk sequence $\{X_n\}$ of the nine species in Table 1.

	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla
$H(X_n^{RY})$	0.445	0.5024	0.6536	0.5075	0.5016	0.538	0.429	0.5791	0.4698
$H(X_n^{MK})$	0.7452	0.7853	0.6547	0.7212	0.7487	0.7094	0.8099	0.5237	0.7467
$H(X_n^{WS})$	0.641	0.6894	0.6292	0.5756	0.6753	0.8118	0.615	0.7255	0.6302

3. Numerical Characterization of DNA Sequences

Researchers from computer science and mathematics have been attracted to study the comparison of DNA sequences. As pointed out in references [13, 16–28], some related work has made progress.

Now, we may represent a DNA sequence by a random numerical sequence based on CGR-walk technique. Gao and Xu [29] also substantially corroborated the results that long-range correlations are uncovered remarkably in the data. In this paper, we explore the tendency of a series of data by calculating the hurst exponent [30]. And some work has been done to study the relation between long-range

correlation and hurst exponent [31]. In order to numerically characterize a DNA sequence given by the CGR, we treat the hurst exponent as the efficient invariant that is sensitive to this kind of graphical representation.

Because a DNA sequence can be regarded as an ordered set of alphabet $\mathcal{N} = (A, C, G, T)$, we represent a DNA sequence as a finite set with N elements, denoted as $[i] := \{1, 2, \dots, N\}$. For any time series $\{u_i\}_{i=1}^N$, one can define several quantities as follows [30]:

(i) the partial mean

$$\langle u \rangle_n = \frac{1}{n} \sum_{i=1}^n u_i, \quad 2 \leq n \leq N, \quad (4)$$

TABLE 3: Mean square deviations of the CGR-walk sequence $\{X_n\}$ of the nine species of in Table 1.

	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla
$D(X_n^{RY})$	0.3979	0.3927	0.3998	0.4192	0.4054	0.3866	0.4060	0.4266	0.3921
$D(X_n^{MK})$	0.3858	0.3949	0.3500	0.3940	0.3636	0.3871	0.3866	0.3908	0.3838
$D(X_n^{WS})$	0.3590	0.3724	0.3907	0.3411	0.4010	0.3912	0.3742	0.3713	0.3574

TABLE 4: Similarity/dissimilarity table for the nine DNA sequences in Table 1 based on Euclidean distance between the 3-component vectors in Table 2.

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla
Human	0	0.0851	0.2277	0.0936	0.0663	0.1978	0.0715	0.2724	0.0271
Goat		0	0.2087	0.1307	0.0392	0.1484	0.1074	0.2750	0.0778
Opossum			0	0.1692	0.1846	0.2229	0.2734	0.1788	0.2056
Gallus				0	0.1036	0.2385	0.1248	0.2581	0.0711
Lemur					0	0.1467	0.1125	0.2432	0.0552
Mouse						0	0.2464	0.2089	0.1976
Rabbit							0	0.3416	0.0767
Rat								0	0.2660
Gorilla									0

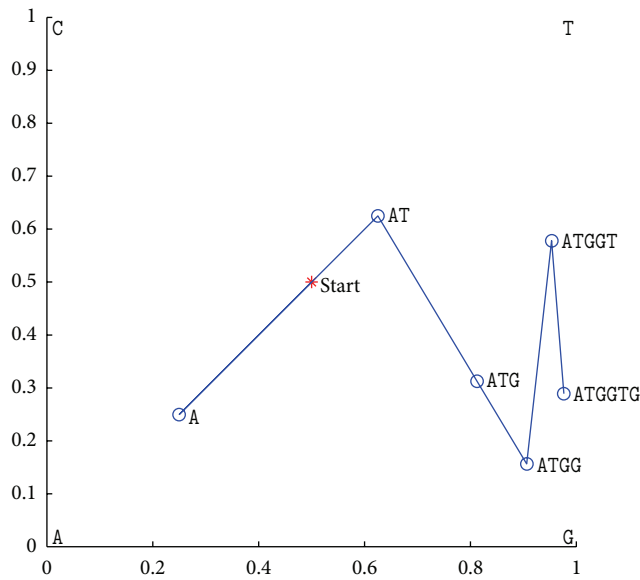


FIGURE 3: CGR-WS of the first 6 bases of exon-1 of human β -globin: ATGGTG.

(ii) the partial difference

$$u(i, n) = \sum_{i=1}^n (u_i - \langle u \rangle_n), \quad 2 \leq n \leq N, \quad (5)$$

(iii) the difference

$$R(n) = \max_{1 \leq i \leq n} \{u(i, n)\} - \min_{1 \leq i \leq n} \{u(i, n)\}, \quad 2 \leq n \leq N, \quad (6)$$

(iv) and the standard deviation

$$S(n) = \left[\frac{1}{n} \sum_{i=1}^n (u_i - \langle u \rangle_n)^2 \right]^{1/2}, \quad 2 \leq n \leq N. \quad (7)$$

Hurst exponent is found to obey the relation:

$$\frac{R(n)}{S(n)} \sim \left(\frac{n}{2}\right)^H, \quad (8)$$

where H is called the hurst exponent.

So we can compute the hurst exponent of RY-, MK- and WS-CGR-walk sequences and characterize the coding sequences of the first exon of β -globin gene of the nine species in Table 1. The results are listed in Table 2.

Besides, there are other numerical characterizations of random sequences, such as the mean, variance, mean square deviation, and so on. Here we choose the mean square deviation of CGR-walk sequence as follows:

$$D(X_i^k) = \left[\frac{1}{N} \sum_{i=1}^N (X_i^k - \mu_{X_i^k})^2 \right]^{1/2}. \quad (9)$$

In (9) k means the classification of RY-, MK-, and WS-sequences, and $\mu_{X_i^k}$ is the mean [13]. We then present the mean square deviations of three kinds of the CGR-walk sequences $\{X_j\}$ in Table 3.

4. Similarity and Dissimilarity among the Coding Sequences of the First Exon of β -Globin Gene of Different Nine Species

Here we construct the three-component vectors in this way, whose components, respectively, are values of hurst exponent and mean square deviation. The analysis of similarity/dissimilarity among DNA sequences represented by the three-component vectors is based on the assumption that two DNA sequences are similar if the corresponding vectors point to one direction in the 3D space. Alternatively we can investigate the similarity among the vectors by calculating

TABLE 5: Similarity/dissimilarity table for the nine DNA sequences in Table 1 based on Euclidean distance between the 3-component vectors in Table 3.

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla
Human	0	0.0171	0.0479	0.0290	0.0481	0.0342	0.0173	0.0317	0.0063
Goat		0	0.0490	0.0410	0.0442	0.0212	0.0157	0.0342	0.0187
Opossum			0	0.0691	0.0180	0.0394	0.0407	0.0526	0.0481
Gallus				0	0.0686	0.0602	0.0364	0.0313	0.0333
Lemur					0	0.0317	0.0353	0.0455	0.0499
Mouse						0	0.0258	0.0449	0.0344
Rabbit							0	0.0213	0.0220
Rat								0	0.0380
Gorilla									0

the Euclidean distance between their end points. Apparently, the smaller the Euclidean distance is, the more similar the two corresponding DNA sequences are.

In Tables 4 and 5, we list the values of Euclidean distances between the 3-component vectors separately including Hurst exponent and mean square deviation. We observe that the smallest entry is always the human-gorilla pair. Furthermore, the largest entries are associated with these rows belonging to opossum (the most remote species from the remaining mammals) and gallus (the only nonmammalian representative). We believe that these results are not accidental, and they coincide with other results in [13, 16–28].

5. Conclusion

DNA sequences play an important role in modern biological research because all the information of the hereditary and species evolution is contained in these macromolecules. How to gain more information from these DNA sequences is still a very challenging question. Description, comparison, and similarity analysis of DNA sequences still occupy important positions.

In this paper, we first construct three kinds of CGR spaces according to the elements of the minor diagonal because the four bases can be classified into R-Y, M-K, and W-S according to their chemical structures. Then we describe a DNA sequence by CGR-walk and convert it to a digital sequence. And we outline some efficient invariants of DNA sequences. As an application, we compare the similarity/dissimilarity of exon-1 of β -globin genes for nine species. From the above tables, we can conclude that the results we got are consistent with known evolutionary facts. Therefore, the method proposed in the paper is visual and efficient.

On one hand, our work can be treated as an effective application of CGR. On the other hand, our method is a valid supplement to graphical representation of DNA sequences. In comparison with other graphical representations of biological sequences, our approach has the following advantages.

- (1) Our graphical representation based on CGR considers the chemical structure classification of the nucleotides and thus may provide more biological information.

- (2) It provides a more simple way of viewing, sorting, and comparing various gene structures, even for longer DNA sequences.

- (3) Our graph is more sensitive, so it can numerically characterize the DNA sequences in a more exact way.

Acknowledgments

The authors thank all the anonymous reviewers for their valuable suggestions and support. This research is supported by the National Science Foundation of China under Grants 11071146 and 10921101.

References

- [1] C. K. Peng, S. V. Buldyrev, A. L. Goldberger et al., “Long-range correlations in nucleotide sequences,” *Nature*, vol. 356, no. 6365, pp. 168–170, 1992.
- [2] J. S. Almeida, J. A. Carriço, A. Maretzek, P. A. Noble, and M. Fletcher, “Analysis of genomic sequences by Chaos Game Representation,” *Bioinformatics*, vol. 17, no. 5, pp. 429–437, 2001.
- [3] H. J. Jeffrey, “Chaos game representation of gene structure,” *Nucleic Acids Research*, vol. 18, no. 8, pp. 2163–2170, 1990.
- [4] S. V. Buldyrev, N. V. Dokholyan, A. L. Goldberger et al., “Analysis of DNA sequences using methods of statistical physics,” *Physica A*, vol. 249, no. 1–4, pp. 430–438, 1998.
- [5] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, H. E. Stanley, and G. M. Viswanathan, *Fractals in Biology and Medicine: from DNA To the Heartbeat*, Springer, Berlin, Germany, 1994.
- [6] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.K. Peng, M. Simons, and H. E. Stanley, “Generalized Lévy-walk model for DNA nucleotide sequences,” *Physical Eview E*, vol. 47, no. 6, pp. 4514–4523, 1993.
- [7] S. V. Buldyrev, A. L. Goldberger, S. Havlin et al., “Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis,” *Physical Review E*, vol. 51, no. 5, pp. 5084–5091, 1995.
- [8] G. Dodin, P. Vanderghenst, P. Levoir, C. Cordier, and L. Marcourt, “Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences,” *Journal of Theoretical Biology*, vol. 206, no. 3, pp. 323–326, 2000.
- [9] A. A. Tsonis, P. Kumar, and J. B. Elsner et al., “Navelet analysis of DNA sequences,” *Physical Review E*, vol. 53, pp. 1828–1834, 1996.

- [10] L. F. Luo, L. Tsai, and Y. M. Zhou, "Informational parameters of nucleic acid and molecular evolution," *Journal of Theoretical Biology*, vol. 130, no. 3, pp. 351–361, 1988.
- [11] L. F. Luo and L. Tsai, "Fractal dimension of nucleic acid and its relation to evolutionary level," *Chemical Physics Letters*, vol. 5, pp. 421–424, 1988.
- [12] A. Arneodo, Y. D'Aubenton-Carafa, E. Bacry, P. V. Graves, J. F. Muzy, and C. Thermes, "Wavelet based fractal analysis of DNA sequences," *Physica D*, vol. 96, no. 1–4, pp. 291–320, 1996.
- [13] F.-L. Bai, Y.-Z. Liu, and T.-M. Wang, "A representation of DNA primary sequences by random walk," *Mathematical Biosciences*, vol. 209, no. 1, pp. 282–291, 2007.
- [14] E. Hamori and J. Ruskin, "H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences," *The Journal of Biological Chemistry*, vol. 258, no. 2, pp. 1318–1327, 1983.
- [15] R. Zhang and C. T. Zhang, "Z-curve, an intuitive tool for visualizing and analyzing the DNA sequences," *Journal of Biomolecular Structure & Dynamics*, vol. 11, pp. 767–782, 1994.
- [16] X. F. Guo, M. Randic, and S. C. Basak, "A novel 2-D graphical representation of DNA sequences of low degeneracy," *Chemical Physics Letters*, vol. 350, no. 1-2, pp. 106–112, 2001.
- [17] M. Randic, "Graphical representations of DNA as 2-D map," *Chemical Physics Letters*, vol. 386, pp. 468–471, 2004.
- [18] G. H. Huang, B. Liao, Y. F. Liu, and Z. B. Liu, "HCL curve: a novel 2D graphical representation for DNA sequences," *Chemical Physics Letters*, vol. 462, pp. 129–132, 2008.
- [19] A. Nandy and P. Nandy, "On the uniqueness of quantitative DNA difference descriptions in 2D graphical representation models," *Chemical Physics Letters*, vol. 368, no. 1-2, pp. 102–107, 2003.
- [20] M. Randic, M. Vracko, N. Lers, and D. Plavsic, "Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation," *Chemical Physics Letters*, vol. 371, pp. 202–207, 2003.
- [21] Y. Yao and T. Wang, "A class of new 2-D graphical representation of DNA sequences and their application," *Chemical Physics Letters*, vol. 398, pp. 318–323, 2004.
- [22] B. Liao and K. Ding, "A 3D graphical representation of DNA sequences and its application," *Theoretical Computer Science*, vol. 358, no. 1, pp. 56–64, 2006.
- [23] Z. Cao, B. Liao, and R. Li, "A group of 3D graphical representation of DNA sequences based on dual nucleotides," *International Journal of Quantum Chemistry*, vol. 108, no. 9, pp. 1485–1490, 2008.
- [24] Y. Huang and T. Wang, "New graphical representation of a DNA sequence based on the ordered dinucleotides and its application to sequence analysis," *International Journal of Quantum Chemistry*, vol. 112, pp. 1746–1757, 2012.
- [25] B. Liao, Y. Zhang, K. Ding, and T. M. Wang, "Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation," *Journal of Molecular Structure*, vol. 717, no. 1–3, pp. 199–203, 2005.
- [26] R. Chi and K. Ding, "Novel 4D numerical representation of DNA sequences," *Chemical Physics Letters*, vol. 407, no. 1-3, pp. 63–67, 2005.
- [27] B. Liao, R. Li, W. Zhu, and X. Xiang, "On the similarity of DNA primary sequences based on 5-D representation," *Journal of Mathematical Chemistry*, vol. 42, no. 1, pp. 47–57, 2007.
- [28] B. Liao and T. M. Wang, "Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 5, pp. 1666–1670, 2004.
- [29] J. Gao and Z. Y. Xu, "Chaos game representation (CGR)-walk model for DNA sequences," *Chinese Physics B*, vol. 18, no. 1, pp. 370–376, 2009.
- [30] Z. G. Yu and V. Anh, "Time series model based on global structure of complete genome," *Chaos, Solitons and Fractals*, vol. 12, no. 10, pp. 1827–1834, 2001.
- [31] L. L. Jiang, Z. Y. Xu, and J. Gao, "Multifractal hurst analysis of DNA sequence," *China Journal of Bioinformatics*, vol. 7, no. 4, pp. 264–267, 2009.