The Scientific World Journal
Volume 2012, Article ID 104269, 6 pages
doi:10.1100/2012/104269

The ScientificWorldJOURNAL

Research Article

Numerical Characterization of DNA Sequence Based on Dinucleotides

Xingqin Qi,[1] Edgar Fuller,[2] Qin Wu,[3] and Cun-Quan Zhang[2]

[1] *School of Mathematics and Statistics, Shandong University at Weihai, Weihai 264209, China*
[2] *Department of Mathematics, West Virginia University, Morgantown, WV 26506, USA*
[3] *School of IOT Engineering, Jiangnan University, Wuxi 214122, China*

Correspondence should be addressed to Xingqin Qi, qixingqin@163.com

Received 4 November 2011; Accepted 26 December 2011

Academic Editors: S. Cacchione and A. Pask

Sequence comparison is a primary technique for the analysis of DNA sequences. In order to make quantitative comparisons, one devises mathematical descriptors that capture the essence of the base composition and distribution of the sequence. Alignment methods and graphical techniques (where each sequence is represented by a curve in high-dimension Euclidean space) have been used popularly for a long time. In this contribution we will introduce a new nongraphical and nonalignment approach based on the frequencies of the dinucleotide XY in DNA sequences. The most important feature of this method is that it not only identifies adjacent XY pairs but also nonadjacent XY ones where X and Y are separated by some number of nucleotides. This methodology preserves information in DNA sequence that is ignored by other methods. We test our method on the coding regions of exon-1 of β–globin for 11 species, and the utility of this new method is demonstrated.

1. Introduction

The number of identifiable DNA sequences responsible for various physiological structures is rapidly increasing as more and more collected DNA sequences are added to scientific databases. It is, however, difficult to obtain information directly from sequences since the sheer volume of data is computational demanding. It is one of the challenges for biologists to analyze mathematically the large volume of genomic DNA sequence data. Many schemes have been proposed to numerically characterize DNA sequences.

Sequence alignment has been used as a very powerful tool for comparison of two closely related genomes at the base-by-base nucleotide sequence level. This method relies heavily on the orderings of nucleotides appearing in the sequence. With the divergence of species over time, though, genomic rearrangements and in particular genetic shuffling make sequence alignment unreliable or impossible.

Graphical techniques are another powerful tool for the analysis and visualization of DNA sequences. Using graphical approaches can provide intuitive pictures or useful insights that assist the analysis of complicated relations between DNA sequences. This methodology starts with a graphical representation of DNA sequence which could be based on 2D, 3D, 4D, 5D, and 6D spaces and represents DNA as matrices by associating with the selected geometrical objects, then vectors composed of the invariants of matrices will be used to compare DNA sequences, see [1–10]. Such schemes have an advantage in that they offer an instant, though, visual and qualitative summary of the lengthy DNA sequences. This approach also involves many unresolved questions. For example, how does one obtain suitable matrices to characterize DNA sequences and how are invariants selected suitable for sequence comparisons? In many cases, the calculation of the matrices or the invariants will become more and more difficult with the length of the sequence. There are also approaches which could arrive a mathematical representation of DNA sequences by nongraphical ways, see [11–13]. And more recently, a new representation based on symbolic dynamics [14] and a new representation based on digital signal method [15] are also illustrated.

In this contribution, we introduce a novel nongraphical and nonalignment approach for DNA sequence comparison. We use DNA sequence directly by considering the frequencies

of dinucleotide. We represent each DNA sequence by a dinucleotide frequency matrix or by a dinucleotide frequency vector, based on which two distance measurements are defined, respectively. Then comparisons between DNA sequences could be carried out by calculating the distances between these mathematical descriptors. The most important feature of this method is that the mathematical descriptors not only take into consideration the frequencies of adjacent XY pairs but also of nonadjacent XY pairs. In this way, information contained in the relative spacing of nucleotides is preserved. The method is very simple and fast, and does not require sequence alignment or sequence graphical representation which would cause complex calculations. It can be used to analyze both short and long DNA sequences. As an application, this method is tested on the exon-1 coding sequences of β-globin for 11 species and the results are consistent with what have been reported previously [5, 9, 12, 14, 15], which prove the utility of this new method.

2. Dinucleotide Frequency Matrix and Dinucleotide Frequency Vector

Typically, DNA sequence data is represented as a string of letters A, C, G, and T, which signify the four nucleotides: adenine, cytosine, guanine, and thymine, respectively. There are 16 possible dinucleotides, that is, $\Omega = \{\text{AT, AA, AC, AG, TT, TA, TC, TG, GT, GA, GC, GG, CT, CA, CC, CG}\}$. In the following, we always use XY to represent dinucleotides, and note that dinucleotide XY is distinguished from.

Let s be a sequence of length n and denote the number of occurrences of adjacent XY in s by $Y^{(1)}$. Clearly, if s is a sequence of length, then $\sum_{XY\in\Omega} XY^{(1)} = n-1$. The occurrence frequency for XY is defined as

$$f_{XY}^{(1)} = \frac{XY^{(1)}}{(n-1)}. \tag{1}$$

We get one 16-dimensional vector $\hat{f}^{(1)}$ associated with sequence s based on adjacent dinucleotides:

$$\hat{f}^{(1)} = \left(f_{\text{AT}}^{(1)}, f_{\text{AA}}^{(1)}, f_{\text{AC}}^{(1)}, \ldots, f_{\text{CT}}^{(1)}, f_{\text{CA}}^{(1)}, f_{\text{CC}}^{(1)}, f_{\text{CG}}^{(1)}\right). \tag{2}$$

Notice that there would be a loss of information when one condenses sequence s to a single 16-dimensional vector. A way to recover some of the lost information associated with a sequence s to a single 16-vector is to introduce additional 16 vectors to store the frequency information of pairs XY when X and Y are not adjacent but are separated at various distance. For example, if s = ATCGATC, the *adjacent* dinucleotides are AT, TC, CG, GA with occurrence frequency 2/6, 2/6, 1/6, and 1/6, respectively. The dinucleotides *at distance* 2 (i.e., separated by one nucleotide) in s are AC, TG, CA, GT, AC with occurrence frequency 2/5, 1/5, 1/5, and 1/5, respectively. These two 16-dimensional vectors will contain additional information beyond that found in the initial dinucleotide vector.

Generally, let s be a sequence of length. Denote $XY^{(d)}$ as the number of occurrence of XY in s when X and Y are separated by $d-1$ nucleotides. Clearly, $\sum_{XY\in\Omega} XY^{(d)} = n-d$. Define

$$f_{XY}^{(d)} = \frac{XY^{(d)}}{(n-d)}, \tag{3}$$

as the occurrence frequency. For each given integer, we could get one 16-dimensional vector $\hat{f}^{(d)}$ associated with sequence s:

$$\hat{f}^{(d)} = \left(f_{\text{AT}}^{(d)}, f_{\text{AA}}^{(d)}, f_{\text{AC}}^{(d)}, \ldots, f_{\text{CT}}^{(d)}, f_{\text{CA}}^{(d)}, f_{\text{CC}}^{(d)}, f_{\text{CG}}^{(d)}\right). \tag{4}$$

The distance d between X and Y could be 1, 2 or even larger integers. When we scan sequence s to count the occurrence of dinucleotides XY at distance, the nucleotides of s from position 1 to $(n-d)$ are counted as "X", while the nucleotides of s from position $(d+1)$ to n are counted as "Y". When $d \le \lfloor (n-1)/2 \rfloor$, there is an overlapping interval $[d+1, n-d]$ between the two intervals $[1, n-d]$ and $[d+1, n]$, which means the nucleotides in the overlapping interval will counted as both X and Y; but if $d > \lceil (n-1)/2 \rceil$, the two intervals $[1, n-d]$ and $[d+1, n]$ will disjoint, and the information of these nucleotides in the interval $[n-d+1, d]$ will be lost. So in the following, to avoid loss of information, d must not be larger than $\lceil (n-1)/2 \rceil$, that is, $d \le \lceil (n-1)/2 \rceil$. Furthermore, to make the information in $\hat{f}^{(d)}$ more accurate, we hope that the overlapping interval $[d+1, n-d]$ will be large enough. Based on this intuition, we would prefer to these d such that $(n-2d)/n \ge 50\%$, which guarantees that more than half of the nucleotides in sequence s will be counted as both X and Y. So d is restricted to $d \le \lfloor n/4 \rfloor$ for each DNA sequence s with length.

Let s be a DNA sequence of length, for a given $d \le \lfloor n/4 \rfloor$, the *dinucleotide frequency matrix* associated with s is defined as

$$F(s) = \begin{pmatrix} \hat{f}^{(1)} \\ \hat{f}^{(2)} \\ \hat{f}^{(3)} \\ \vdots \\ \hat{f}^{(d)} \end{pmatrix}, \tag{5}$$

where $\hat{f}^{(i)}$ is the 16-dimensional occurrence frequency vector when X and Y are separated by $(i-1)$ nucleotides. The size of matrix $F(s)$ is $d \times 16$.

We also present another mathematical descriptor associated with s named *dinucleotide frequency vector* which is defined as

$$\hat{F}(s) = \left(\hat{f}^{(1)}, \hat{f}^{(2)}, \hat{f}^{(3)}, \ldots, \hat{f}^{(d)}\right), \tag{6}$$

then $\hat{F}(s)$ is a $1 \times 16d$ row vector.

3. Two Distance Measurements Based on Dinucleotide Frequency

From Section 2, we get correspondences between one DNA sequence s and the dinucleotide frequency matrix $F(s)$ and

the dinucleotide frequency vector $\hat{F}(s)$. Note that the sizes of $F(s)$ and $\hat{F}(s)$ all depend on. To make the comparisons for a set of DNA sequences meaningful, we should use an identical d for all these DNA sequences. Denote the set of DNA sequences by, by the discussion in Section 2, we define the identical d_0 as

$$d_0 = \min_{s \in S} \left\lfloor \frac{(|s|)}{4} \right\rfloor, \tag{7}$$

where $|s|$ is the length of s. The choice of d_0 will guarantee that either the frequency matrix or the frequency vector will involve enough accurate information, and the dinucleotide frequency matrices and dinucleotide frequency vectors associated with sequences in S all have the same size. DNA sequences comparisons could be completed by studying their corresponding matrices and vectors. In the following we will introduce two different distance measurements based on dinucleotide frequencies matrix and dinucleotide frequency vector, respectively.

3.1. City Block Distance for Dinucleotide Frequency Matrix. Given two DNA sequences s and h, then we get the dinucleotide frequency matrix $F(s)$ and $F(h)$ as in Section 2, comparison between s and h becomes comparison between $F(s)$ and $F(h)$. Using this, we define the city block distance $d_1(s,h)$ between s and h as

$$d_1(s,h) = \sum_{1 \le i \le d_0,\ 1 \le j \le 16} \left| F_{ij}(s) - F_{ij}(h) \right|. \tag{8}$$

3.2. Cosine Distance for Dinucleotide Frequency Vector. We also obtain a mapping from a DNA sequence s to a vector $\hat{F}(s)$ in the $16d_0$-dimensional linear space. Comparison between DNA sequences also could become comparison between these $16d_0$-dimensional vectors. This is based on the assumption that two DNA sequences are similar if the corresponding $16d_0$-dimensional vectors in the $16d_0$-dimensional space have similar directions. Given two DNA sequences s and h, the dinucleotide frequency vectors are $\hat{F}(s)$ and $\hat{F}(h)$, we define the cosine distance $d_2(s,h)$ between s and h as

$$d_2(s,h) = 1 - \cos\left(\hat{F}(s), \hat{F}(h)\right), \tag{9}$$

where $\cos(\hat{F}(s), \hat{F}(h))$ is the cosine value of the included angle between vectors $\hat{F}(s)$ and $\hat{F}(h)$.

4. Applications and Experimental Results

4.1. Experimental Results. A comparison between a pair of DNA sequences to judge their similarity or dissimilarity could be carried out by calculating the distance $d_1(s,h)$ or $d_2(s,h)$. The smaller is the distance, the much similar are the two DNA sequences (The code is available on request).

To test the utility of above method, we make a comparison for the coding regions of exon-1 of β-globin gene for 11 different species, which were also studied by Randić et al. in [12]. Table 1 presents their accession numbers in NCBI database, while Table 2 lists these 11 coding sequences concretely.

Table 1: ID Information for Exon-1 of β -globin gene of 11 species.

Species	ID/Accession	Database	length
Human	U01317	NCBI	92
Chimpanzee	X02345	NCBI	105
Gorilla	X61109	NCBI	93
Lemur	M15734	NCBI	92
Rat	X06701	NCBI	92
Mouse	V00722	NCBI	93
Rabbit	V00882	NCBI	92
Goat	M15387	NCBI	86
Bovine	X00376	NCBI	86
Opossum	J03643	NCBI	92
Gallus	V00409	NCBI	92

At first, we present the similarity/dissimilarity matrix based on distance measurement d_1, see Table 3. When we examine this table, we notice that smallest entries are always associated with the pairs (human, chimpanzee) with $d_1 = 2.5567$, (human, gorilla) with $d_1 = 2.4026$, and (gorilla, chimpanzee) with $d_1 = 2.7338$. That means the more similar species pairs are human-gorilla, human-chimpanzee, and gorilla-chimpanzee. We also observe that the largest entry $d_1 = 9.0347$ is associated with gallus and lemur and the larger entries appear in the rows belonging to gallus and opossum, which is consistent with the facts that gallus is the only nonmammalian species among these 11 species and opossum is the most remote species from the remaining mammals. These observed facts are consistent with the results reported in previous studies [5, 9, 12] determined by matrix invariants techniques, and also consistent with the reported results from nongraphical means [14, 15]. More interesting, in Table 3, the distance between goat and bovine is $d_1 = 2.3438$, which is actually the smallest entry in Table 3. That implies goat and bovine are regarded to be much similar to each other by our method, which is consistent with their biology taxonomy that bovine and goat are both even-toed ungulates and belong to the family of "Bovidae".

Table 4 presents the similarity/dissimilarity matrix based on the distance measurement d_2. The smallest entries are also associated with the pairs (human, chimpanzee) with $d_2 = 0.0087$, (human, gorilla), with $d_2 = 0.0074$, and (gorilla, chimpanzee), and with $d_2 = 0.0112$. We find that the largest entry ($d_2 = 0.1139$) is associated with (gallus, lemur), and the rows corresponding to gallus and opossum have larger entries, which is also consistent with the facts that gallus is the only nonmammalian species among these 11 species and opossum is the most remote species from the remaining mammals. The observed facts in Table 4 are consistent with the previously reported results in [5, 9, 12, 14, 15] as well. And the distance between goat and bovine ($d_2 = 0.0109$) is also much smaller as we expect.

We can see that there is an overall qualitative agreement between Tables 3 and 4. To see it visually, we denote

Table 2: The coding sequence of exon-1 of β -globin gene for 11 species.

Species	DNA sequence
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG

Table 3: The upper triangular part of the dissimilarity/similarity matrix based on d_1.

Species	Human	Chimpanzee	Gorilla	Lemur	Rat	Mouse	Rabbit	Goat	Bovine	Opossum	Gallus
Human	0	2.5567	2.4026	6.4922	5.6622	4.9144	4.2904	5.3220	4.8306	6.8358	7.4959
Chimpanzee		0	2.7338	6.5340	5.9455	5.1613	4.9587	5.6525	4.9670	7.4568	7.9791
Gorilla			0	7.0466	6.2344	5.2819	5.0310	5.3353	4.9340	7.8956	8.0582
Lemur				0	6.9735	6.8419	5.6647	6.9332	6.0195	8.2293	9.0347
Rat					0	5.2540	6.8004	6.5847	6.2545	7.5359	8.2347
Mouse						0	6.5730	6.7863	6.4133	7.2900	7.8317
Rabbit							0	5.9265	5.2974	8.0743	8.3210
Goat								0	2.3438	8.0158	7.7129
Bovine									0	7.9847	8.2938
Opossum										0	8.0268
Gallus											0

Table 4: The upper triangular part of the dissimilarity/similarity matrix based on d_2.

Species	Human	Chimpanzee	Gorilla	Lemur	Rat	Mouse	Rabbit	Goat	Bovine	Opossum	Gallus
Human	0	0.0087	0.0074	0.0567	0.0464	0.0372	0.0253	0.0354	0.0287	0.0719	0.0819
Chimpanzee		0	0.0112	0.0564	0.0487	0.0383	0.0303	0.0403	0.0320	0.0793	0.0899
Gorilla			0	0.0619	0.0538	0.0398	0.0312	0.0357	0.0302	0.0887	0.0877
Lemur				0	0.0691	0.0635	0.0454	0.0616	0.0463	0.0939	0.1139
Rat					0	0.0417	0.0631	0.0592	0.0552	0.0832	0.1048
Mouse						0	0.0588	0.0573	0.0528	0.0765	0.0932
Rabbit							0	0.0444	0.0349	0.0998	0.0933
Goat								0	0.0109	0.0948	0.0792
Bovine									0	0.0923	0.0937
Opossum										0	0.0897
Gallus											0

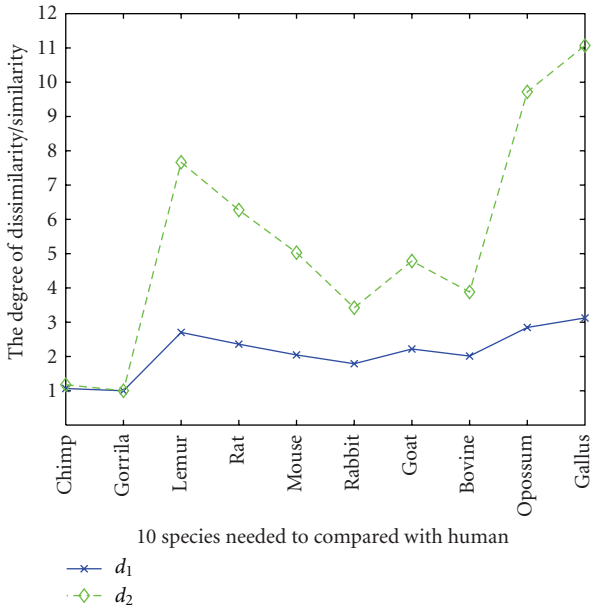

Figure 1: The degree of dissimilarity/similarity of the other 10 species with human, where the degree of dissimilarity/similarity of the pair human-gorilla is defined relatively as 1.

the degree of dissimilarity/similarity of the pair human-gorilla as 1 in each table, then the results of the examination of the degree of dissimilarity/similarity between human and other several species under the two distance measurements are shown in Figure 1. We can see that the curvilinear trend of these two curves are almost the same, which demonstrates the overall agreement among dissimilarity/similarities obtained by these two distance methods.

4.2. Discussion. For the above exon-1 coding data of 11 species, d_0 is chosen to be 21 followed by (7). A 336-dimensional vector is used to characterize each DNA sequence under the second distance measure. To confirm the efficacy of the vectors constructed in this high-dimensional data representation, we perform principal component analysis (PCA) on these 336 parameters. Figure 2(a) shows the projection of the 11 vectors on a 2D property space composed of the top two principal components PC1, PC2. We can see that in the 2D space, gallus (labeled by "⊙") and opossum (labeled by "∇") are furthest from the other 9 species, and human, chimpanzee, and gorilla are very close to each other. These result are consistent with what we have got from Table 4. Note that these top two principal components contribute 48% (see Figure 2(b)) to the total information. Some information is lost when we do the projection, for example, bovine seems much closer to rabbit than goat in the 2D projection, but we know this is not true in Table 4 when all 336 parameters are considered. However, this rough approximation confirms that our mathematical descriptor characterizes DNA sequence structure effectively.

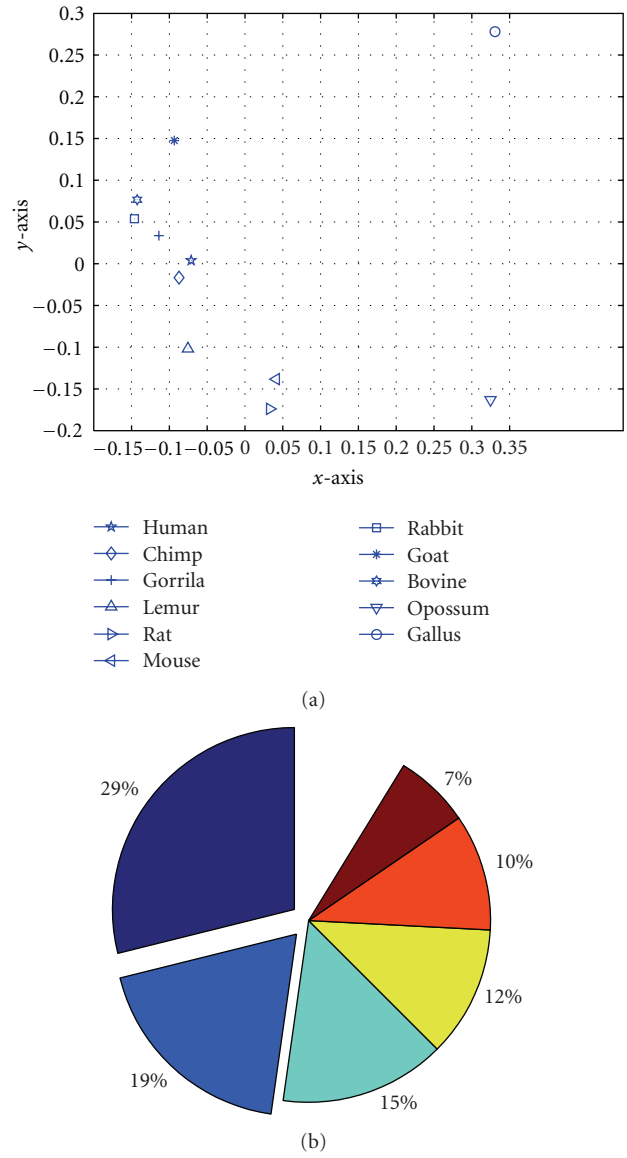

Figure 2: (a) The projection of the 336-dimensional vectors of 11 species on a 2D space composed of the top two principal components; (b) The contributions of the first 6 principal components.

5. Conclusion

In this paper, we have presented a new method based on dinucleotide frequencies for DNA sequence comparison. The dinucleotide frequency matrix and dinucleotide frequency vector are used to mathematically characterize a DNA sequence. The most important feature of this method is that the mathematical descriptors not only involve the frequencies of adjacent XY pairs but also nonadjacent XY pairs (i.e., when X and Y are separated by various number of nucleotides), such that a lot of important information is avoided to lose. This new method does not require sequence alignment or sequence graphical representation, which avoids the complex calculation found in either sequence alignment or sequence graphical representation.

The method is very simple and fast, and it can be used to analyze both short and long DNA sequences with high efficiencies.

Acknowledgments

This work is supported partly by Shandong Province Natural Science Foundation of China with no. ZR2010AQ018 and no. ZR2011FQ010 and partly by Independent Innovation Foundation of Shandong University with no. 2010ZRJQ005. This project also has been partially supported by a WV EPSCoR Grant and an NSA Grant H98230-12-1-0233.

References

[1] E. Hamori and J. Ruskin, "H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences," *Journal of Biological Chemistry*, vol. 258, no. 2, pp. 1318–1327, 1983.

[2] A. Nandy, "A new graphical representation and analysis of DNA sequence structure: I. Methodology and Application to Globin Genes," *Current Science*, vol. 66, pp. 309–314, 1994.

[3] M. Randić, M. Vračko, A. Nandy, and S. C. Basak, "On 3-D graphical representation of DNA primary sequences and their numerical characterization," *Journal of Chemical Information and Computer Sciences*, vol. 40, no. 5, pp. 1235–1244, 2000.

[4] Y. Zhang, B. Liao, and K. Ding, "On 2D graphical representation of DNA sequence of nondegeneracy," *Chemical Physics Letters*, vol. 411, no. 1-3, pp. 28–32, 2005.

[5] M. Randić, M. Vračko, N. Lerš, and D. Plavšić, "Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation," *Chemical Physics Letters*, vol. 371, no. 1-2, pp. 202–207, 2003.

[6] B. Liao and T. M. Wang, "3-D graphical representation of DNA sequences and their numerical characterization," *Journal of Molecular Structure (THEOCHEM)*, vol. 681, no. 1–3, pp. 209–212, 2004.

[7] Y. Zhang, B. Liao, and K. Ding, "On 3DD-curves of DNA sequences," *Molecular Simulation*, vol. 32, no. 1, pp. 29–34, 2006.

[8] R. Chi and K. Ding, "Novel 4D numerical representation of DNA sequences," *Chemical Physics Letters*, vol. 407, no. 1–3, pp. 63–67, 2005.

[9] 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.

[10] 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.

[11] M. Randić, "Condensed representation of DNA primary sequences," *Journal of Chemical Information and Computer Sciences*, vol. 40, no. 1, pp. 50–56, 2000.

[12] M. Randić, X. Guo, and S. C. Basak, "On the characterization of DNA primary sequences by triplet of nucleic acid bases," *Journal of Chemical Information and Computer Sciences*, vol. 41, no. 3, pp. 619–626, 2001.

[13] Y. Zhang, "A simple method to construct the similarity matrices of DNA sequences," *Match*, vol. 60, no. 2, pp. 313–324, 2008.

[14] S. Wang, F. Tian, W. Feng, and X. Liu, "Applications of representation method for DNA sequences based on symbolic dynamics," *Journal of Molecular Structure: THEOCHEM*, vol. 909, no. 1–3, pp. 33–42, 2009.

[15] Z. H. Qi and X. Q. Qi, "Numerical characterization of DNA sequences based on digital signal method," *Computers in Biology and Medicine*, vol. 39, no. 4, pp. 388–391, 2009.