**MATCH**
*Communications in Mathematical and in Computer Chemistry*

MATCH Commun. Math. Comput. Chem. **61** (2009) 781-788

ISSN 0340 - 6253

# Three distances for rapid similarity analysis of DNA sequences

Wei Chen, Yusen Zhang*
*School of Mathematics and Statistics, Shandong University at Weihai*
*Weihai 264209, China*

*(Received July 14, 2008)*

**Abstract.** Three distances for assessing genomic similarity based on dinucleotide frequency in large DNA sequences is introduced. The method requires neither homologous sequences nor prior sequence alignments. The analysis centers on symmetrized dinucleotide frequency reflecting DNA structures related to dinucleotide stacking energies, constraints of DNA curvature. To show the utility of the method, we use these distances to examine the similarities among the first exon-1 of the $\beta$-globin gene for 11 different species.

## 1 Introduction

The traditional algorithms for similarity analysis and phylogenetic inference are based mostly on multiple alignment [1]. Such approaches have been hitherto widely used. However, for large genomic sequences, alignments of the sequences are generally not feasible. To overcome these problems, more and more researchers begin to try alignment-free methods for DNA sequence comparison and analysis [2, 3]. More recently, alternative routes for quantitative measure of DNA sequences were considered [4,5]. The novel methodology starts with a graphical representation of DNA, such as proposed in [6-10], which allow visual inspection of lengthy sequences. It was shown that it is possible to characterize numerically the graphical representation to obtain a numerical characterization of the degree of similarity /dissimilarity of different DNA sequences. This is accomplished by associating with graphical representations of DNA a corresponding mathematical object such as a matrix, and then using various properties of mathematical object, like matrix invariants, as sequence descriptors. In this way one arrives at an alternative approach for comparative studies of DNA, which are less computer-intensive, because it replaces the original DNA sequence by an ordered set of sequence invariants, which can be viewed as components of vectors and thus comparison of sequences is transformed into a simpler comparison of vectors, rather than by a direct comparison of the sequences themselves [11-16]. An important advantage of the characterization of structures by this invariants, as opposed to use of codes, is the simplicity

*Corresponding author: zhangys@sdu.edu.cn

of the comparison based on invariants. Although we also use other invariants [17][18], the calculation of invariants, especially the eigenvalues, will become more and more difficult with the order of the matrix large. In paper [19], a new method based on the double-stranded nature of DNA has been proposed to construct the similarity matrices. Such matrix allows one to mathematically characterize the DNA sequences and make quantitative comparisons between different DNA sequences, between the same or between different species.

In this paper we consider the properties of the neighboring dual nucleotides of DNA sequence and propose three distances based on symmetrized dinucleotide frequency reflecting DNA structures related to dinucleotide stacking energies, constraints of DNA curvature., which are adaptive to both analysis of short and long DNA sequences. As an application, we make a comparison for the first exon-1 of the $\beta$-globin gene for 11 different species.

## 2 Symmetrized dinucleotide frequency

Consider a DNA sequence read from the 5'- to the 3'-end with $n$ bases. The cumulative numbers of the nucleotide X (A, C, G, or T), denoted by the positive integer $X_n$. By considering neighboring two bases, we can obtain sixteen dinucleotide XY: AG, GA, CT, TC, AC, CA, GT, TG, AT, TA, CG, GC, AA, CC, GG and TT. The cumulative numbers of the dinucleotide XY denoted by the positive integer $XY_n$. Let $f_X$ denote the frequency of the nucleotide X (A, C, G, or T) and $f_{XY}$ denote the frequency of dinucleotide XY. Then we obtain $f_X = X_n/n$ and $f_{XY} = XY_n/n - 1$. Since DNA structures are influenced by oligonucleotide compositions of both strands (e.g., stacking energies), the frequency formula for $f_{XY}$ is modified to accommodate the double-stranded nature of DNA by combining the given sequence and its inverted complement sequence. In this context, the frequency $f_A$ is symmetrized to $f'_A = f'_T = (f_A + f_T)/2$ and $f'_C = f'_G = (f_C + f_G)/2$. Similarly, $f'_{GT} = f'_{AC} = (f_{GT} + f_{AC})/2$ is the symmetrized double stranded frequency of GT/AC, etc.

## 3 Proposed distance

Before we present our main result, we define three distances between two DNA sequences.

Let f is a DNA sequence, the dinucleotide frequency matrix is defined by:

$$F(f) = \begin{bmatrix} f'_{AT} & f'_{AA} & f'_{AC} & f'_{AG} \\ f'_{TT} & f'_{TA} & f'_{TC} & f'_{TG} \\ f'_{GT} & f'_{GA} & f'_{GC} & f'_{GG} \\ f'_{CT} & f'_{CA} & f'_{CC} & f'_{CG} \end{bmatrix}. \tag{1}$$

By this way, we get a correspondence between the DNA sequence and the dinucleotide frequency matrix $F(f)$. A DNA sequence can be analyzed by studying the corresponding dinucleotide frequency matrix. It is easy to find that the dinucleotide frequency matrix $F(f)$ is real symmetric.

Given two sequences $f$ and $g$ (e.g., sequences from different organisms or from different regions of a genomic sequence), then we get the dinucleotide frequency matrix $F(f)$ and

$F(g)$, comparison between DNA sequences becomes comparison between these dinucleotide frequency matrices. Based on this idea the first dinucleotide frequency distance $d_1(f,g)$ is defined as

$$d_1(f,g) = \sum |F_{ij}(f) - F_{ij}(g)|,$$

where the sum extends over all dinucleotides.

Another distance measure that would follow from the idea of building the variance-covariance matrix corresponding to the dinucleotide frequency matrix $F(f)$.

The variance-covariance matrix $CF(f)$ consists of the variances of the variables along the main diagonal and the covariances between each pair of variables in the other matrix positions. If the row vectors of dinucleotide frequency matrix $F(f)$ are denoted as $X_i$ $(i = 1, 2, 3, 4)$, then the formula for computing the covariance of $X_i$ and $X_j$ is

$$CF(f)_{ij} = \frac{1}{4}\sum_{k=1}^{4}(X_i - X_i^0)(X_j - X_j^0),$$

where $X_i^0$ and $X_j^0$ denoting the means of $X_i$ and $X_j$, respectively.

Given two sequences $f$ and $g$, the second distance measure is defined as

$$d_2(f,g) = \sum w_{ij}|CF(f)_{ij} - CF(g)_{ij}|,$$

where the sum extends over all dinucleotides and $w_{ij} = 16$ or some other natural weights.

Table 1: The coding sequences of the first exon of $\beta$-globin gene of eleven different species

| Species | Coding sequence |
|---|---|
| human | ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGT GAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG |
| Goat | ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAAGT GGATGAAGTTGGTGCTGAGGCCCTGGGCAG |
| Opossum | ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGT GCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG |
| Gallus | ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGT CAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG |
| Lemmur | ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGT GGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG |
| Mouse | ATGGTGCACCTGACTGATGCTGAGAAGGCTGCTGTCTCTTGCCTGTGGGGAAAGGT GAACTCCGATGAAGTTGGTGGTGAGGCCCTGGGCAG |
| Rabbit | ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGT GATTGTGGAAGAAGTTGGTGGTGAGGCCCTGGGCAG |
| Rat | ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGT GAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG |
| Gorilla | ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGT GAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG |
| Bovine | ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAAAGT GGATGAAGTTGGTGGTGAGGCCCTGGGCAG |
| Chimpanzee | ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGT GAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG |

Before we present the third distance, we need to consider ratio measure $P_{XY}$, that is the ratio of the frequency $f_{XY}$ of a dinucleotide XY to the frequency $f_X$ of the nucleotide X. The ratio measure $P_{XY}$ is suitable for a single sequence, in order to compare sequences

from different organisms (or from different chromosomes), the formula has to be modified to account for the complementary antiparallel structure of double-stranded DNA. Based on this fact, the $P_{XY}$ is defined as $P_{XY} = f'_{XY}/f'_X$. That is, $P_{GT} = f'_{GT}/f'_G = n(f_{GT} + f_{AC})/((n-1)(f_C + f_G))$, and similarly for other dinucleotides. For any DNA sequence f, we construct a 16-component vector:

$$V(f) = (P_{AA}, P_{AT}, P_{AG}, P_{AC}, P_{AT}, P_{TT}, P_{TG}, P_{TC}, P_{AG}, P_{TG}, P_{GG}, P_{GC}, P_{AC}, P_{TC}, P_{GC}, P_{CC}),$$

then we get a correspondence between the DNA sequence and 16-component vector $V(f)$. So Given two sequences $f$ and $g$, the third distance measure can be defined as

$$d_3(f,g) = \sqrt{\sum (P_{ij}(f) - P_{ij}(g))^2},$$

where the sum extends over all dinucleotides.

A comparison between a pair of DNA sequences to judge their similarities and dissimilarities can be carried out by calculatting the distance $d_3(f,g)$. The analysis of similarity among these DNA sequences is based on the assumption that the smaller is the Euclidean distance the more similar are the two DNA sequences.

Table 2: The upper triangular part of the similarities/dissimilarities matrix based on distance measure $d_1(f,g)$ of the 11 coding sequences of Table 1

| *Species* | Human | Goat | Opossum | Gallus | Lemur | Mouse | Rabbit | Rat | Gorilla | Bovine | Chimpanzee |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Human | 0 | 0.0732 | 0.1464 | 0.0895 | 0.1018 | 0.0833 | 0.0725 | 0.0502 | 0.0329 | 0.0493 | 0.0249 |
| Goat | | 0 | 0.1161 | 0.0605 | 0.0849 | 0.0836 | 0.0902 | 0.0814 | 0.0583 | 0.0408 | 0.0862 |
| Opossum | | | 0 | 0.1547 | 0.0523 | 0.0981 | 0.1083 | 0.1356 | 0.1350 | 0.1419 | 0.1490 |
| Gallus | | | | 0 | 0.1082 | 0.0900 | 0.0965 | 0.0902 | 0.0761 | 0.0591 | 0.0978 |
| Lemur | | | | | 0 | 0.0723 | 0.0904 | 0.0870 | 0.0959 | 0.0900 | 0.1004 |
| Mouse | | | | | | 0 | 0.0283 | 0.0808 | 0.0754 | 0.0747 | 0.0864 |
| Rabbit | | | | | | | 0 | 0.0932 | 0.00570 | 0.0819 | 0.0714 |
| Rat | | | | | | | | 0 | 0.0685 | 0.0462 | 0.0557 |
| Gorilla | | | | | | | | | 0 | 0.0537 | 0.0350 |
| Bovine | | | | | | | | | | 0 | 0.0522 |
| Chimpanzee | | | | | | | | | | | 0 |

## 4 Results and discussion

The computation of the proposed distances is simple and alignment-free. Unlike most existing methods to analyze the similarity of DNA sequence, the proposed method does not require gene identification nor any prior biology knowledge such as an accurate alignment score matrix. To show the utility of the method, we use these distances to examine the similarities among the first exon-1 of the $\beta$-globin gene for 11 different species. In Table 1, the first

exon-1 of the $\beta$-globin gene for 11 different species are listed, which were reported by Randić et al. [7].

Table 2 presents the similarities/dissimilarities matrix based on distance measure $d_1(f, g)$ of the 11 coding sequences of Table 1. The smallest entries are associated with the pairs human and Chimpanzee [$d_1 = 0.0249$], mouse and rabbit [$d_1 = 0.0283$], human and gorilla [$d_1 = 0.0329$] and gorilla and chimpanzee [$d_1 = 0.0350$]. The greatest distance, $d_1 = 0.1547$, among the 11 coding sequences is observed between gallus (the only non-mammalian representative) and opossum (the most remote species from the remaining mammals), and the larger entries in the similarity matrix appear in the rows belonging to opossum and gallus.

Table 3: The upper triangular part of the similarities/dissimilarities matrix based on distance measure $d_2(f, g)$ of the 11 coding sequences of Table 1

| *Species* | Human | Goat | Opossum | Gallus | Lemur | Mouse | Rabbit | Rat | Gorilla | Bovine | Chimpanzee |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Human | 0 | 0.2061 | 0.1978 | 0.2198 | 0.2637 | 0.1676 | 0.1292 | 0.1758 | 0.0545 | 0.1590 | 0.0467 |
| Goat | | 0 | 0.2534 | 0.1500 | 0.1766 | 0.1515 | 0.2157 | 0.1950 | 0.2128 | 0.0941 | 0.2276 |
| Opossum | | | 0 | 0.3516 | 0.1319 | 0.1728 | 0.1771 | 0.2637 | 0.2131 | 0.2976 | 0.1841 |
| Gallus | | | | 0 | 0.2637 | 0.2110 | 0.1995 | 0.2418 | 0.1710 | 0.1580 | 0.2170 |
| Lemur | | | | | 0 | 0.1531 | 0.2225 | 0.1978 | 0.2554 | 0.2097 | 0.2610 |
| Mouse | | | | | | 0 | 0.0865 | 0.1560 | 0.1590 | 0.1442 | 0.1648 |
| Rabbit | | | | | | | 0 | 0.2279 | 0.0982 | 0.1922 | 0.1108 |
| Rat | | | | | | | | 0 | 0.2267 | 0.1606 | 0.2033 |
| Gorilla | | | | | | | | | 0 | 0.1714 | 0.0602 |
| Bovine | | | | | | | | | | 0 | 0.1805 |
| Chimpanzee | | | | | | | | | | | 0 |

In Table 3, the similarities/dissimilarities matrix is based on distance measure $d_2(f, g)$. Observing Table 3, we find that gallus is very dissimilar to others among the 11 species because its corresponding row has lager entries. which is consistent with the fact that Gallus is non-mammal, while the others are mammal. On the other hand, the two close species are human and chimpanzee [$d_2(human, chimpanzee) = 0.0467$], human and gorilla [$d_2(human, gorilla) = 0.0545$], and gorilla and chimpanzee [$d_2(chimpanzee, gorilla) = 0.0602$], the distance [$d_2 = 0.0865$] between mouse and rabbit, dissimilar with that in Table 2, is larger than above mentioned three cases.

Comparing Table 2, 3, we can find that there exists an overall qualitative agreement among similarities although there is small difference. The result presented in Table 2 and Table 3 are in accord with that reported results of the examination of the similarity/dissimilarity of the coding sequences of the first exon of $\beta$-globin gene of several species by means of approaches using matrix invariants techniques[12][14][15][17][18].

We present the the similarities/dissimilarities matrix obtained using the distance $d_3(f, g)$ in Table 4. The similarities is in agreement with the results in Table 2 and Table 3 except for the greatest distance which are associated with gallus and chimpanzee [$d_3(chimpanzee, gallus) =$

Table 4: The upper triangular part of the similarities/dissimilarities matrix based on distance measure $d_3(f, g)$ of the 11 coding sequences of Table 1

| *Species* | Human | Goat | Opossum | Gallus | Lemur | Mouse | Rabbit | Rat | Gorilla | Bovine | Chimpanzee |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Human | 0 | 0.3225 | 0.1666 | 0.3612 | 0.2943 | 0.1677 | 0.1602 | 0.2558 | 0.0835 | 0.2417 | 0.0741 |
| Goat | | 0 | 0.2877 | 0.2287 | 0.1956 | 0.2211 | 0.3178 | 0.2815 | 0.3273 | 0.1527 | 0.3604 |
| Opossum | | | 0 | 0.3292 | 0.1923 | 0.1535 | 0.1719 | 0.2770 | 0.1708 | 0.2576 | 0.1745 |
| Gallus | | | | 0 | 0.2297 | 0.2291 | 0.2771 | 0.2749 | 0.3452 | 0.2676 | 0.3621 |
| Lemur | | | | | 0 | 0.1809 | 0.2604 | 0.2410 | 0.3057 | 0.2053 | 0.3010 |
| Mouse | | | | | | 0 | 0.1193 | 0.2240 | 0.1633 | 0.1618 | 0.1737 |
| Rabbit | | | | | | | 0 | 0.2942 | 0.1135 | 0.2687 | 0.1325 |
| Rat | | | | | | | | 0 | 0.3062 | 0.2079 | 0.2708 |
| Gorilla | | | | | | | | | 0 | 0.2717 | 0.0962 |
| Bovine | | | | | | | | | | 0 | 0.2745 |
| Chimpanzee | | | | | | | | | | | 0 |

0.3621], not gallus and opossum. the two close species are (human, Chimpanzee) [$d_3 = 0.0741$], (human, gorilla) [$d_3 = 0.0835$] and (gorilla, chimpanzee) [$d_3 = 0.0962$]. But Gallus is dissimilar to others in evidence. And we can clearly verify that gallus and opossum are dissimilar to others in Table 4. Besides gallus and opossum, lemur should be more remote from the other species relatively.

## 5 Conclusions

Sequence comparison is a fundamental task in Computational Biology that aims to discover similarity relationships between molecular sequences. In this paper, three distances based on the symmetrized dinucleotide frequency of DNA sequence has been proposed to mathematically characterize the DNA sequences. Its application to the similarity/dissimilarity of the coding sequences of $\beta$-globin gene of 11 species and each of the exons of the gene illustrate validity. The results about similarity fix basically the reality. Meanwhile, our approach does not require complicated calculation. The method is more simple, convenient and fast. So they are adaptive to both analysis of short and long DNA sequences.

## Acknowledgements

This work was supported in part by the Shandong Natural Science Foundation(Y2006A14).

## References

[1] A. Godzik, The structural alignment between two proteins: is there a unique answer? *Protein Sci.* 5 (1996) 1325-1338.

[2] C. Burge, A. M. Campbell, and S. Karlin., Over- and under-representation of short oligonucleotides in DNA sequences, *Proc. Natl. Acad. Sci.* 89 (1992) 1358-1362.

[3] S. Karlin, I. Ladunga, Comparisons of eukaryotic genomic sequences, *Proc. Natl. Acad. Sci.* 91(1994), 12832-12836,

[4] A. Nandy, M. Harle, S.C. Basak, Mathematical descriptors of DNA sequences: development and applications, *ARKIVOC* 9 (2006) 211-238.

[5] G. Jaklic, T. Pisanski, M. Randić, Characterization of complex biological systems by matrix invariants, *J. Comput. Biol.* 13 (2006) 1558-1564.

[6] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Novel 2-D graphical representation of DNA sequences and their numberical characterization, *Chem. Phys. Lett.* 368 (2003) 1-6.

[7] M. Randić, X. F. Guo, S. C. Basak, On the characterization of DNA primary sequence by triplet of nucleic acid bases, *J. Chem. Inf. Comput. Sci.* 41(2001) 619-626.

[8] Y. S. Zhang, B. Liao, K. Ding, On 2D graphical representation of DNA sequence of nondegeneracy, *Chem. Phys. Lett.* 411 (2005) 28-32.

[9] Y. S. Zhang, B. Liao, K. Ding, On 3DD-Curves of DNA sequences, *Mol. Simul.* 32(2006) 29-34.

[10] B. Liao, A 2D graphical representation of DNA sequence,*Chem. Phys. Lett.* 401(2005) 196-199.

[11] Y. S. Zhang, M. Tan, Visualization of DNA sequences based on 3DD-Curves, *J. Math. Chem.* 44 (2008) 206-216.

[12] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, *Chem. Phys. Lett.* 371(2003) 202-207.

[13] D. Bielinska-Waz, P. Waz, T. Clark, Similarity studies of DNA sequences using genetic methods, *Chem. Phys. Lett.* 445 (2007) 68-73.

[14] B. Liao, Y. S. Zhang, K. Ding, T. Wang, Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation, *J. Mol. Struct. (Theochem)* 717 (2005) 199-203.

[15] B. Liao, K. Ding, Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleaotide bases, *J. Comput. Inf. Comput. Sci.* 44 (2004) 1666-1670.

[16] B. Liao, K. Ding, A graphical approach to analyzing DNA sequences, *J. Comput. Chem.* 26 (2005) 1519-1523.

[17] Y. S. Zhang, W. Chen, Invariants of DNA sequences based on 2DD-curves, *J. Theor. Biol.* 242 (2006) 382-388.

[18] Y. S. Zhang, W. Chen , New invariant of DNA sequences, *MATCH Commun. Math. Comput. Chem.* 58 (2007), 207-218.

[19] Y. S. Zhang, A simple method to construct the similarity matrices of DNA sequences, *MATCH Commun. Math. Comput. Chem.* 60 (2008) 313-324.