MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **61** (2009) 533-540

ISSN 0340 - 6253

Comparisons of DNA Sequences Based on Dinucleotide

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

(Received May 5, 2008)

Abstract. In this paper a method for assessing DNA similarity based on dinucleotide frequencies in DNA sequence is introduced. The method does not require prior sequence alignments. The analysis centers on dinucleotide frequencies in DNA sequences and distances between sequences based on Frobenius norm of covariance matrices of dinucleotide frequencies. Analysis shows an overall qualitative agreement among similarities for the beta globin exon 1 sequences of 11 species.

1 Introduction

The rapid accumulation of large numbers of DNA sequences affords challenging opportunities for studies of molecular evolution and phylogenetic relationships among organisms. The primary structure of DNA consists basically of a nitrogenous base of four nucleotides, the two purines, adenine (A) and guanine (G), and the two pyrimidines, cytosine (C) and thymine (T). Thus the DNA sequence can be simply considered as a symbolic sequence on the four symbols A,C,G and T. For a long time the computer science approach was the only methodology. There is a family of methods which relies on initial alignment of homologous DNA or protein sequences followed by tree construction based on various principles, which is based on assuming particular scoring functions, that introduce various penalties for the existence of insertions or deletions in the alignment. As has been described by Godzik [1], the outcome of such searches need not be unique. More recently, alternative routes for

*Corresponding author: zhangys@sdu.edu.cn

quantitative measure of the degree of similarity of DNA sequences were considered [2, 3]. The novel methodology starts with a graphical representation of DNA, such as proposed in [4, 8, 9, 10, 13], which are subsequently numerically characterized by associating with the selected geometrical object that represents DNA a matrix [5, 7, 11]. For example, one can consider distance matrix in which matrix elements are given as the distances between the vertices which form the geometrical representation of the sequence. Alternatively, one can consider the quotients of distances measured through space and measure along the shortest path between pairs of vertices [2, 7, 11]. Finally, we should add that one can arrive at a matrix representation of DNA sequence also without graphical representation. One such representation is based on using the overall sequential labels and sequential labels of each of the four nucleotides A, T, G, and C separately for construction of matrix elements [5]. Construction of matrices to represent DNA has an important advantage for characterization of DNA in that instead of direct comparison of sequences one can construct vectors, the components of which are various matrix invariants. The similarity between sequences is then transformed in calculation of similarities between n-dimensional vectors [12].

In this paper we consider the double-stranded nature of DNA and utility the double stranded frequency of dinucleotides to make similarities analysis among DNA sequences. We also propose a new numerical representation of DNA by matrices the elements of which indicate some measure between the dinucleotides measured by the frequency of dinucleotides in DNA sequences. As an application, we make a comparison for the first exon-1 of the β-globin gene for 11 different species. In Table 1, the first exon-1 of the β-globin gene for 11 different species are listed, which were reported by Randić et al.[5, 6].

2 Dinucleotide symmetrized frequencies

Let us consider a DNA sequence read from the 5'- to the 3'-end with n bases. 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 in the subsequence from the first to the i-th base in the DNA sequence is denoted by the positive integer XY_i. Let f_i^{XY} denote the

Table 1: The coding sequences of the first exon of β-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

frequency of dinucleotide XY, occurring in the subsequence from the first to the i-th base in the DNA sequence, where $i = 1, 2, ..., n-1$. Then we obtain $f_i^{XY} = XY_i/i$.

Since DNA structures are influenced by oligonucleotide compositions of both strands (e.g., stacking energies), the frequency formula for f_i^{XY} is modified to accommodate the double-stranded nature of DNA by combining the given sequence and its inverted complement sequence. In this context, considering the base pairs of A and T, and of C and G, the symmetrized double stranded frequency of XY is defined as follows:

$$
\begin{aligned}
f_i^1 &= f_i^{TA} = f_i^{TA}, \\
f_i^2 &= f_i^{AT} = f_i^{AT}, \\
f_i^3 &= f_i^{CG} = f_i^{CG}, \\
f_i^4 &= f_i^{GC} = f_i^{GC}, \\
f_i^5 &= f_i^{GT} = f_i^{AC} = (f_i^{GT} + f_i^{AC})/2, \\
f_i^6 &= f_i^{AG} = f_i^{CT} = (f_i^{AG} + f_i^{CT})/2, \\
f_i^7 &= f_i^{CA} = f_i^{TG} = (f_i^{CA} + f_i^{TG})/2, \\
f_i^8 &= f_i^{TC} = f_i^{GA} = (f_i^{TC} + f_i^{GA})/2, \\
f_i^9 &= f_i^{GG} = f_i^{CC} = (f_i^{GG} + f_i^{CC})/2, \\
f_i^{10} &= f_i^{AA} = f_i^{TT} = (f_i^{AA} + f_i^{TT})/2,
\end{aligned}
$$

where, $i = 1, 2, ..., n-1$.

In Table 2, we list the total dinucleotide symmetrized double stranded frequency XY occurring in the 11 coding sequences of Table 1.

Table 2: The symmetrized frequency XY in the 11 coding sequences of Table 1

Species	$f_{n-1}^{AA}\%$	$f_{n-1}^{TA}\%$	$f_{n-1}^{CG}\%$	$f_{n-1}^{GC}\%$	$f_{n-1}^{AT}\%$	$f_{n-1}^{TC}\%$	$f_{n-1}^{GG}\%$	$f_{n-1}^{CA}\%$	$f_{n-1}^{AG}\%$	$f_{n-1}^{GT}\%$
Human	3.8462	2.1978	2.1978	6.5934	2.1978	4.9451	10.4396	9.3407	7.6923	7.1429
Goat	4.1176	0	2.3529	10.5882	2.3529	5.8824	9.4118	9.4118	9.4118	4.1176
Opossum	3.8462	2.1978	0	5.4945	3.2967	6.5934	7.1429	10.4396	9.3407	7.1429
Gallus	2.7473	0	3.2967	10.9890	4.3956	5.4945	10.9890	9.8901	7.6923	3.8462
Lemur	4.3956	1.0989	1.0989	7.6923	4.3956	7.1429	7.1429	9.8901	9.3407	4.9451
Mouse	3.8462	0	1.0989	7.6923	3.2967	6.5934	8.7912	10.4396	8.7912	5.4945
Rabbit	3.2967	0	1.0989	6.5934	3.2967	6.0440	9.8901	10.9890	7.6923	6.5934
Rat	4.3956	4.3956	1.0989	8.7912	4.3956	3.8462	9.3407	9.3407	8.2418	5.4945
Gorilla	3.2609	1.0870	2.1739	6.5217	2.1739	5.4348	10.8696	9.7826	7.6087	7.0652
Bovine	5.2941	0	2.3529	9.4118	2.3529	5.2941	10.0000	9.4118	8.2353	4.7059
Chimpanzee	3.8462	1.9231	1.9231	5.7692	2.8846	5.2885	10.5769	9.6154	7.2115	7.2115

3 Comparison of DNA sequences

3.1 Comparison based on 10-component frequency vector

We illustrate the use of the symmetrized frequencies of DNA sequences with the examination of similarities/dissimilarities among the 11 coding sequences of Table 1. For any DNA sequence g, we construct a 10-component frequency vector which is defined as:

$$V(g) = (f_{n-1}^1, f_{n-1}^2, f_{n-1}^3, f_{n-1}^4, f_{n-1}^5, f_{n-1}^6, f_{n-1}^7, f_{n-1}^8, f_{n-1}^9, f_{n-1}^{10}),$$

where n is the number of bases making up the corresponding DNA sequence. Then we get a correspondence between the DNA sequence and 10-component vectors $V(g)$.

A comparison between a pair of DNA sequences to judge their similarities and dissimilarities can be carried out by comparing their associated frequency vectors, which is equivalent to calculate the distance $d(f, g)$, which is a measure of symmetrized frequency distance between two sequences f and g, and is defined to be the Euclidean distance between the end points of 10-component vectors $V(f)$ and $V(g)$.

The analysis of similarity among these DNA sequences represented by the 10-component vectors is based on the assumption that two DNA sequences are similar

if the corresponding 10-component vectors have similar magnitudes. The similarity between these two vectors can be measured by calculating the Euclidean distance between their end points. Clearly, the smaller is the Euclidean distance the more similar are the two DNA sequences.

Table 3: The upper triangular part of the similarities/dissimilarities matrix based on the Euclidean distances between the end points of the 10-component vectors of the 11 coding sequences of Table 1

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	0	0.0591	0.0498	0.0657	0.0568	0.0436	0.0348	0.0463	0.149	0.0463	0.0132
Goat		0	0.0729	0.0361	0.0480	0.0389	0.0556	0.0598	0.0574	0.0228	0.0661
Opossum			0	0.0884	0.0337	0.04407	0.0433	0.0599	0.0522	0.0684	0.0477
Gallus				0	0.0641	0.0533	0.0599	0.0636	0.0616	0.0404	0.0694
Lemur					0	0.0338	0.0450	0.0544	0.0575	0.0488	0.0571
Mouse						0	0.0239	0.0560	0.0391	0.0356	0.0439
Rabbit							0	0.0599	0.00257	0.0462	0.0301
Rat								0	0.0560	0.0542	0.0512
Gorilla									0	0.0454	0.0158
Bovine										0	0.0522
Chimpanzee											0

Table 3 presents all Euclidean distances between the end points of the 10-component vectors of sequences of Table 1. The greatest distance, $d = 0.0884$, among the 11 coding sequences is observed between gallus (the only non-mammalian representative) and opossum (the most remote species from the remaining mammals), gallus is not particularly close to others. The two close species are human and chimpanzee [$d(human, chimpanzee) = 0.0132$], human and gorilla [$d(human, gorilla) = 0.0149$], and gorilla and chimpanzee [$d(chimpanzee, gorilla) = 0.0158$].

3.2 Comparison based on covariance matrix

For any DNA sequence g, the mathematical expectation of f_i^k (f_i^{XY}), denoted by f_0^k (f_0^{XY}), may be calculated as follows:

$$f_0^k = \frac{1}{n-1} \sum_{i=1}^{n-1} f_i^k.$$

we can construct a 10×10 covariance matrix

$$CM(g) = (CM_{ij}(g)),$$

where,

$$CM_{ij}(g) = \frac{1}{n-1}\sum_{k=1}^{n-1}(f_k^i - f_0^i)(f_k^j - f_0^j),$$

and $i, j = 1, 2, ..., 10$.

Then we get a correspondence between the DNA sequence and the covariance matrix $CM(g)$. A DNA sequence can be analyzed by studying the corresponding covariance matrix. Comparison between sequences becomes comparison between these covariance matrices.

A direct comparison of DNA sequences using the cumulative numbers of the dinucleotide XY will yield poor comparative distance assessments due to the fact that the DNA sequence have different lengths, so we use the dinucleotide symmetrized frequency of XY (f_i^{XY}) instead of the cumulative numbers of the dinucleotide XY (XY_i) to construct covariance matrix.

Table 4: The upper triangular part of the similarities/dissimilarities matrix based on dinucleotide frequency distance $\delta(f, g)$ for the 11 coding sequences

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	0	0.0139	0.0136	0.0174	0.0149	0.0131	0.0094	0.0117	0.0029	0.0124	0.0062
Goat		0	0.0204	0.0083	0.0154	0.0091	0.0119	0.0177	0.0127	0.0049	0.0148
Opossum			0	0.0226	0.0087	0.0116	0.0122	0.0137	0.0139	0.0213	0.0150
Gallus				0	0.0161	0.0127	0.0130	0.0193	0.0151	0.0102	0.0172
Lemur					0	0.0066	0.0105	0.0131	0.0135	0.0167	0.0144
Mouse						0	0.0063	0.0131	0.0115	0.0103	0.0128
Rabbit							0	0.0140	0.0068	0.0120	0.0094
Rat								0	0.0138	0.0165	0.0138
Gorilla									0	0.0113	0.0052
Bovine										0	0.0136
Chimpanzee											0

In order to compare the DNA sequences by using the covariance matrix of DNA sequence, we introduce the dinucleotide frequency distance $\delta(f, g)$, which is a measure of dinucleotide distance between two sequences f and g, and is defined as

$$\delta(f, g) = \sum_{i,j} |CM_{ij}(f) - CM_{ij}(g)|,$$

where the sum extends over all dinucleotides.

The dinucleotide frequency distance $\delta(f,g)$ of covariance matrix can be used as mathematical descriptors of the DNA sequences to quantitatively compare the sequences and determine similarities and dissimilarities between them. Comparison of two DNA sequences is then transformed in calculation of $\delta(f,g)$ of the corresponding DNA sequences. Clearly, the smaller is the distance the more similar are the two DNA sequences.

The similarities/dissimilarities matrix based on dinucleotide frequency distance $\delta(f,g)$ for the 11 coding sequences of Table 1 is listed in Table 4.

From Table 4, we find that the smallest entries are associated with the pairs human and gorilla [$\delta = 0.0029$], gorilla and chimpanzee [$\delta = 0.0052$] and human and Chimpanzee [$\delta = 0.0062$]. And we further verify that gallus and opossum are dissimilar to others. Besides gallus and opossum, lemur should be more remote from the other species relatively. The main results are similar to that reported in previous studies.

4 Conclusions

The new method based on the double-stranded nature of DNA has been proposed to mathematically characterize the DNA sequences. Such matrix allows one to make quantitative comparisons between different DNA sequences, between the same or between different species. Unlike the other methods of making analysis of similarity/dissimilarity, the new method does not require sequence alignment. The methods utilize the entire information contained in the DNA sequences and avoid the complex calculation as in the calculation of invariants in matrices of higher order. 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). The authors would like to thank the anonymous referees and editors for their corrections and valuable comments.

References

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

[2] 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.

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

[4] 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.

[5] 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.

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

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

[8] B. Liao, K. Ding, A graphical approach to analyzing DNAsequences, *J. Comput. Chem.*, **26** (2005) 1519–1523.

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

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

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

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

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