Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Journal of Theoretical Biology 266 (2010) 703–707

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

New method for comparing DNA primary sequences based on a discrimination measure

Jie Feng[a,b,*], Yong Hu[c], Ping Wan[c], Aibing Zhang[c], Weizhong Zhao[a,b]

[a] *School of Mathematical Sciences, Capital Normal University, Beijing 100048, China*
[b] *Institute of Mathematics and Interdisciplinary Science, Capital Normal University, Beijing 100048, China*
[c] *College of Life Sciences, Capital Normal University, Beijing 100048, China*

ARTICLE INFO

Article history:
Received 24 March 2010
Received in revised form
23 June 2010
Accepted 28 July 2010
Available online 3 August 2010

Keywords:
Pairwise distance
Similarity analysis
Phylogenetic tree

ABSTRACT

We introduce a new approach to compare DNA primary sequences. The core of our method is a new measure of pairwise distances among sequences. Using the primitive discrimination substrings of sequence S and Q, a discrimination measure $DM(S,Q)$ is defined for the similarity analysis of them. The proposed method does not require multiple alignments and is fully automatic. To illustrate its utility, we construct phylogenetic trees on two independent data sets. The results indicate that the method is efficient and powerful.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

With the completion of the sequencing of the genomes of human and other species, the field of analysis of genomic sequences is becoming very important tasks in bioinformatics. Comparison of primary sequences of different DNA strands remains the upmost important aspect of the sequence analysis. So far, most comparison methods are based on string alignment (Pearson and Lipman, 1988; Lake, 1994): a distance function is used to represent insertion, deletion, and substitution of letters in the compared strings. Using the distance function, one can compare DNA primary sequences and resolve the questions of the homology of macromolecules. However, it is not easy to use for long sequences since it is realized with the aid of dynamic programming, which will be slow due to the large number of computational steps.

In the past two decades, alignment-free sequence comparison (Vinga and Almeida, 2003) has been actively pursued. Some new methods have been derived with a variety of theoretical foundations. One category out of these methods is based on the statistics of word frequency within a DNA sequence (Sitnikova and Zharkikh, 1993; Karlin and Burge, 1995; Wu et al., 1997, 2001; Stuart et al., 2002; Qi et al., 2004). The core idea is that the more similar the two sequences are, the greater the number of the factors shared by two sequences is. The earliest publication using frequencies statistics of k-words for sequence comparison dates from 1986 (Blaisdell, 1986). Three years after, Blaisdell (1989) proved that the dissimilarity values observed by using distance measures based on word frequencies are directly related to the ones requiring sequence alignment. In recent years, many researchers employ the k-words and the Markov model to obtain the information about the biological sequences (Pham and Zuegg, 2004; Pham, 2007; Kantorovitz et al., 2007; Helden, 2004; Dai et al., 2008).

Another category does not require resolving the sequence with fixed word length segments. It can be further divided into three groups. In the first group, researchers represent DNA sequence by curves (Hamori and Ruskin, 1983; Nandy, 1994; Randic et al., 2003a; Zhang et al., 2003; Liao, 2005; Li et al., 2006; Qi et al., 2007; Yu et al., 2009), numerical sequences (He and Wang, 2002), or matrices (Randic, 2000; Randic et al., 2001). According to the representation, some numerical characterizations are selected as invariants of sequence for comparisons of DNA primary sequences. The advantage of these methods is that they provide a simple way of viewing, sorting and comparing various gene structures. But how to obtain suitable invariants to characterize DNA sequences and compare them is still a question need our attention.

The second group corresponds to iterated maps. Jeffrey (1990) proposed the chaos game representation (CGR) as a scale-independent representation for genomic sequences. The algorithm exploited iterative function systems to map nucleotide sequences into a continuous space. Since then, alignment-free methods based on CGR have aroused much interest in the field of computational biology. Further studies by Almeida et al. (2001) showed that CGR is a generalized Markov chain probability table which can accommodate non-integer orders, and that CGR

[*] Corresponding author at: School of Mathematical Sciences, Capital Normal University, Beijing 100048, China. Tel.: +86 010 68905511x410.
E-mail address: fengjie0536@163.com (J. Feng).

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.07.040

is a powerful sequence modelling tool because of its computational efficiency and scale-independence (Almeida and Vinga, 2002, 2006, 2009). Such alignment-free methods have been successfully applied for sequence comparison, phylogeny, detection of horizontal transfers, detection of oligonucleotides of interest, meta-genomic studies (Deschavanne et al., 1999; Pride et al., 2003; Sandberg et al., 2003; Teeling et al., 2004; Chapus et al., 2005; Wang et al., 2005; Dufraigne et al., 2005; Joseph and Sasikumar, 2006).

The third group is based on text compression technique (Li et al., 2001; Chen et al., 2004; Cilibrasi et al., 2004). If one sequence which is given the information contained in the other sequence is significantly compressible, the two sequences are considered to be close. There are also some important methods which are based on compression algorithm but do not actually apply the compression, such as Lemple–Ziv complexity and Burrows–Wheeler transform (Otu and Sayood, 2003; Mantaci et al., 2007, 2008; Yang et al., 2010).

In this paper, we propose a new sequence distance for the similarity analysis of DNA sequences. Based on the properties of primitive discrimination substrings, we construct a discrimination measure (DM) between every two sequences. Furthermore, as application, two data sets (β–globin genes and coronavirus genomes) are prepared and tested to identify the validity of the method. The results demonstrate that the new method is powerful and efficient.

2. Discrimination measure

DNA sequences consist of four nucleotides: A (adenine), G (guanine), C (cytosine), and T (thymine). A DNA sequence, of length n, can be viewed as a linear sequence of n symbols from a finite alphabet $\mathcal{A}=\{A,C,G,T\}$. Let S and Q be sequences defined over $\mathcal{A}$, $l(S)$ be the length of S, $S(i)$ denotes the ith element of S and $S(i,j)$ is the substring of S composed of the elements of S between positions i and j (inclusive).

Definition 1. *$S(i, j)$ is called a discrimination substring (DS) that distinguishes S from Q if $S(i,j)\overline{\in}Q$, particularly, if $S(i, j)$ does not include any other DSs distinguishing S from Q, we call $S(i, j)$ a primitive discrimination substring (PDS) that distinguishes S from Q.*

The set of PDSs that distinguish S from Q is denoted by $\Delta(S,Q)$. Similarly, $\Delta(Q,S)$ expresses the set of PDSs that distinguish Q from S. Note that every sequence has its own identity, hence $\Delta(S,Q)$ is usually different from $\Delta(Q,S)$. For example for $S=acctac$ and $Q=gtgact$, we can obtain that $\Delta(S,Q)=\{cc,ta\}$ and $\Delta(Q,S)=\{gt,tg,ga,act\}$.

Suppose $u\in\Delta(S,Q)$ and $l(u)=k$, then we can get $u(1,k-1)\in Q$ (otherwise $u(1,k-1)\in\Delta(S,Q)$, which conflicts with the minimum of u). Therefore the larger the k is, the more the same elements both S and Q have and correspondingly the smaller the degree of discrimination that S distinguishes from Q is. On the other hand, if the number of appearances of u in sequence S is t, we obviously note that the smaller the t is, the smaller the degree of discrimination that S distinguishes from Q is. From the above description, we construct the following discrimination measure that one sequence distinguishes from another sequence.

Definition 2. $DM(S_1\to S_2)$ denotes the discrimination measure that S_1 distinguishes from S_2

$$DM(S\to Q)=\sum_{u\in\Delta(S,Q)} t/[(l(S)-k+1)\log_2(k+1)],$$

$$DM(Q\to S)=\sum_{v\in\Delta(Q,S)} t'/[(l(Q)-k'+1)\log_2(k'+1)],$$

in which $v\in\Delta(Q,S)$, $l(v)=k'$ and the number of appearances of v in sequence Q is t'.

Definition 3. The discrimination measure of sequences S and Q is

$$DM(S,Q)=\sqrt{(DM(S\to Q))^2+(DM(Q\to S))^2}.$$

For the function DM to be a distance, it must satisfy (a) $DM(x,y)>0$ for $x\neq y$; (b) $DM(x,x)=0$; (c) $DM(x,y)=DM(y,x)$ (symmetric); and (d) $DM(x,y)\le DM(x,z)+DM(z,y)$ (triangle inequality). Apparently, DM satisfies distance conditions (a)–(c). It is not obvious that it also satisfies (d). The following proposition answers this.

Proposition 1. *$DM(x,y)$ satisfies the triangle inequality, that is $DM(x,z)\le DM(x,y)+DM(y,z)$.*

Proof. Suppose s is an arbitrary element of $\Delta(x,z)$. If s is also contained in $\Delta(x,y)$, clearly we can obtain that $DM(x\to z)\le DM(x\to y)+DM(y\to z)$. If there exists an element $t\in\Delta(x,z)$, and t is not contained in $\Delta(x,y)$, then we can derive $t\in\Delta(y,z)$, therefore the triangle inequality $DM(x\to z)\le DM(x\to y)+DM(y\to z)$ still comes into existence. Similarly, we can prove that $DM(z\to x)\le DM(y\to x)+DM(z\to y)$.

Let $a=DM(x\to z)$, $b=DM(z\to x)$, $c=DM(x\to y)$, $d=DM(y\to x)$, $e=DM(y\to z)$, $f=DM(z\to y)$, we need to show

$$\sqrt{a^2+b^2}\le\sqrt{c^2+d^2}+\sqrt{e^2+f^2}.$$

Since

$$\sqrt{a^2+b^2}\le\sqrt{(c+e)^2+(d+f)^2},$$

it is sufficient to prove the following inequality:

$$\sqrt{(c+e)^2+(d+f)^2}\le\sqrt{c^2+d^2}+\sqrt{e^2+f^2}.$$

This is equivalent to, by squaring both sides of the above inequality,

$$ce+df\le\sqrt{(c^2+d^2)(e^2+f^2)}.$$

To prove this inequality, we just need to prove

$$(ce+df)^2\le(c^2+d^2)(e^2+f^2),$$

i.e. $2cedf\le e^2d^2+c^2f^2$. Obviously, this inequality comes into existence. Therefore,

$$\sqrt{a^2+b^2}\le\sqrt{c^2+d^2}+\sqrt{e^2+f^2}.$$

Hence $DM(x,y)$ satisfies the triangle inequality. □

3. Results and discussion

In this section, we apply the discrimination measure to analyze two sets of DNA primary sequences. The similarities among these species are computed by calculating the discrimination measure between every two sequences. The smaller the discrimination measure is, the more similar the species are. That is to say, the discrimination measures of evolutionary closely related species are smaller, while those of evolutionary disparate species are larger. Fig. 1 illustrates the basic processes of the DM algorithm. The first set we select includes 10 β–globin genes, whose similarity has been studied by many researchers using their first exon sequences (Randic et al., 2003b; Liu and Wang, 2005). Here we will analyze these species using their complete β–globin genes. Table 1 presents their names, EMBL accession numbers, locations and lengths.

In Table 2, we present the similarity/dissimilarity matrix for the full DNA sequences of β−globin gene from 10 species listed in Table 1 by our new method. Observing Table 2, we note that the most similar species pairs are human–gorilla, human–chimpanzee and gorilla–chimpanzee, which is expected as their evolutionary relationship. At the same time, we find that gallus and opossum are the most remote from the other species, which coincides with the fact that gallus is the only nonmammalian species among these 10 species and opossum is the most remote species from the remaining mammals. By further study of the values in the table, we can gain more information about their similarity.

Another usage of the similarity/dissimilarity matrix is that it can be used to construct phylogenetic tree. The quality of the constructed tree may show whether the matrix is good and therefore whether the method of abstracting information from DNA sequences is efficient. Once a distance matrix has been calculated, it is straightforward to generate a phylogenetic tree using the NJ method or the UPGMA method in the PHYLIP package (http://evolution.genetics.washington.edu/phylip.html). In Fig. 2, we show the phylogenetic tree of 10 β−globin gene sequences based on the distance matrix DM, using NJ method. The tree is drawn using the DRAWGRAM program in the PHYLIP package. From this figure, we observe that (1) gallus is clearly separated from the rest, this coincides with real biological phenomenon; (2) human, gorilla, chimpanzee and lemur are placed closer to bovine and goat than to mouse and rat, this is in complete agreement with Cao et al. (1998) confirming the outgroup status of rodents relative to ferungulates and primates.

Next, we consider inferring the phylogenetic relationships of coronaviruses with the complete coronavirus genomes. The 24 complete coronavirus genomes used in this paper were downloaded from GenBank, of which 12 are SARS-CoVs and 12 are from other groups of coronaviruses. The name, accession number, abbreviation, and genome length for the 24 genomes are listed in Table 3. According to the existing taxonomic groups, sequences 1–3 form group I, and sequences 4–11 belong to group II, while sequence 12 is the only member of group III. Previous work showed that SARS-CoVs (sequences 13–24) are not closely related to any of the previously characterized coronaviruses and form a distinct group IV.

In Fig. 3, we present the phylogenetic tree belonging to 24 species based on the distance matrix DM, using UPGMA method. The tree is viewed using the DRAWGRAM program. As shown in Fig. 3, four groups of coronaviruses can be seen from it: (1) The group I coronaviruses, including TGEV, PEDV and HCoV-229E tend to cluster together; (2) BCoV, BCoVL, BCoVM, BCoVQ, MHV, MHV2, MHVM, and MHVP, which belong to group II, are grouped in a monophyletic clade; (3) IBV, belonging to group III, is situated at an independent branch; (4) the SARS-CoVs from group IV are grouped in a separate branch, which can be distinguished easily

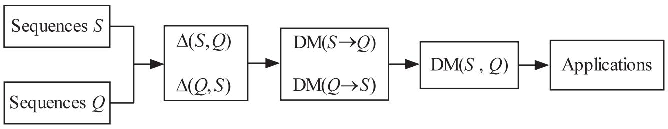

Fig. 1. The flow diagram of our method.

Table 1
The full DNA sequences of β−globin gene of 10 species.

Species	Database	Accession	Location	Length (bp)
Human	EMBL	U01317	62187–63610	1424
Goat	EMBL	M15387	279–1749	1471
Opossum	EMBL	J03643	467–2488	2022
Gallus	EMBL	V00409	465–1810	1346
Lemur	EMBL	M15734	154–1595	1442
Mouse	EMBL	V00722	275–1462	1188
Rat	EMBL	X06701	310–1505	1196
Gorilla	EMBL	X61109	4538–5881	1344
Bovine	EMBL	X00376	278–1741	1464
Chimpanzee	EMBL	X02345	4189–5532	1344

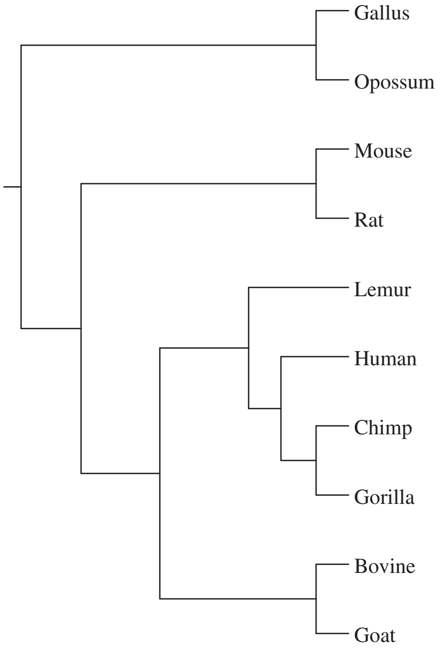

Fig. 2. The phylogenetic tree for 10 species using the full DNA sequences of β−globin gene based on DM.

Table 2
The similarity/dissimilarity matrix for 10 β−globin genes based on DM.

Species	Human	Goat	Oposs.	Gallus	Lemur	Mouse	Rat	Gorilla	Bovine	Chimp.
Human	0	0.2394	0.2683	0.2803	0.2233	0.2484	0.2489	0.0294	0.2438	0.0297
Goat		0	0.2761	0.2844	0.2502	0.2634	0.2570	0.2474	0.1130	0.2471
Oposs.			0	0.2895	0.2828	0.2728	0.2698	0.2701	0.2785	0.2733
Gallus				0	0.2827	0.2778	0.2838	0.2895	0.2805	0.2896
Lemur					0	0.2591	0.2502	0.2223	0.2545	0.2201
Mouse						0	0.1863	0.2548	0.2606	0.2530
Rat							0	0.2562	0.2628	0.2558
Gorilla								0	0.2472	0.0167
Bovine									0	0.2483
Chimp.										0

Table 3
The accession number, abbreviation, name, and length for each of the 24 coronavirus genomes.

No.	Accession	Abbreviation	Genome	Length (nt)
1	NC_002645	HCoV-229E	Human coronavirus 229E	27,317
2	NC_002306	TGEV	Transmissible gastroenteritis virus	28,586
3	NC_003436	PEDV	Porcine epidemic diarrhea virus	28,033
4	U00735	BCoVM	Bovine coronavirus strain Mebus	31,032
5	AF391542	BCoVL	Bovine coronavirus isolate BCoV-LUN	31,028
6	AF220295	BCoVQ	Bovine coronavirus Quebec	31,100
7	NC_003045	BCoV	Bovine coronavirus	31,028
8	AF208067	MHVM	Murine hepatitis virus strain ML-10	31,233
9	AF201929	MHV2	Murine hepatitis virus strain 2	31,276
10	AF208066	MHVP	Murine hepatitis virus strain Penn 97-1	31,112
11	NC_001846	MHV	Murine hepatitis virus	31,357
12	NC_001451	IBV	Avian infectious bronchitis virus	27,608
13	AY278488	BJ01	SARS coronavirus BJ01	29,725
14	AY278741	Urbani	SARS coronavirus Urbani	29,727
15	AY278491	HKU-39849	SARS coronavirus HKU-39849	29,742
16	AY278554	CUHK-W1	SARS coronavirus CUHK-W1	29,736
17	AY282752	CUHK-Su10	SARS coronavirus CUHK-Su10	29,736
18	AY283794	SIN2500	SARS coronavirus Sin2500	29,711
19	AY283795	SIN2677	SARS coronavirus Sin2677	29,705
20	AY283796	SIN2679	SARS coronavirus Sin2679	29,711
21	AY283797	SIN2748	SARS coronavirus Sin2748	29,706
22	AY283798	SIN2774	SARS coronavirus Sin2774	29,711
23	AY291451	TW1	SARS coronavirus TW1	29,729
24	NC_004718	TOR2	SARS coronavirus	29,751

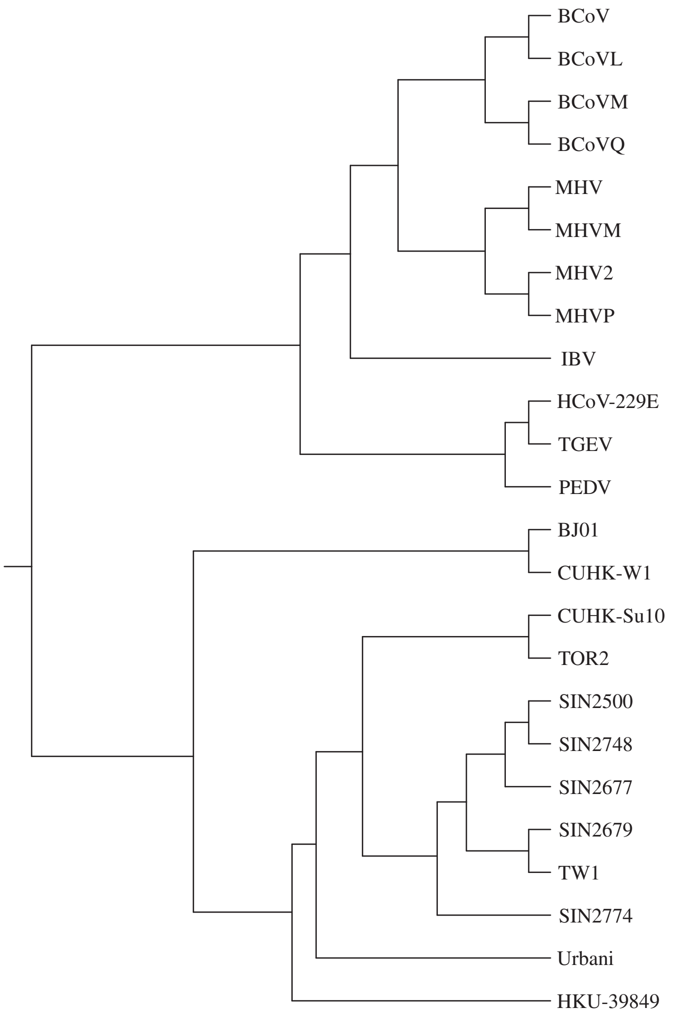

Fig. 3. The phylogenetic tree for 24 coronavirus using whole genomes based on DM.

from other three groups of coronaviruses. The tree constructed based on DM algorithm is quite consistent with the results obtained by other researchers (Zheng et al., 2005; Song et al., 2005; Liu et al., 2007; Li et al., 2008). The emphasis of the present work is to provide a new method to analyze DNA sequences. From the above applications, we can see that our method is feasible for comparing DNA sequences and deducing their similarity relationship.

4. Conclusion

In this paper, we propose a new method for the similarity analysis of DNA sequences. It is a simple method that yields results reasonably and rapidly. Our algorithm is not necessarily an improvement as compared to some existing methods, but an alternative for the similarity analysis of DNA sequences. The new approach does not require sequence alignment and graphical representation, and besides, it is fully automatic. The whole operation process utilizes the entire information contained in the DNA sequences and do not require any human intervention. The application of the DM algorithm to the sets of $\beta-$globin genes and coronavirus genomes demonstrates its utility. This method will also be useful to researchers who are interested in evolutionary analysis.

Acknowledgements

We thank all the anonymous referees for their valuable suggestions and support.

References

Almeida, J.S., Carrico, J.A., Maretzek, A., Noble, P.A., Fletcher, M., 2001. Analysis of genomic sequences by chaos game representation. Bioinformatics 17, 429–437.

Almeida, J.S., Vinga, S., 2002. Universal sequence map (USM) of arbitrary discrete sequences. BMC Bioinformatics 3, 6.

Almeida, J.S., Vinga, S., 2006. Computing distribution of scale independent motifs in biological sequences. Algorithms Mol. Biol. 1, 18.

Almeida, J.S., Vinga, S., 2009. Biological sequences as pictures: a generic two dimensional solution for iterated maps. BMC Bioinformatics 10, 100.

Blaisdell, B., 1986. A measure of similarity of sets of sequences not requiring sequence alignment. Proc. Natl. Acad. Sci. 83, 5155–5159.

Blaisdell, B., 1989. Effectiveness of measures requiring and not requiring prior sequence alignment for estimating the dissimilarities of natural sequences. J. Mol. Evol. 29, 526–537.

Cao, Y., Janke, A., Waddell, P.J., Westerman, M., Takenaka, O., Murata, S., Okada, N., Paabo, S., Hasegawa, M., 1998. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J. Mol. Evol. 47, 307–322.

Chapus, C., Dufraigne, C., Edwards, S., Giron, A., Fertil, B., Deschavanne, P., 2005. Exploration of phylogenetic data using a global sequence analysis method. BMC Evol. Biol. 5, 63.

Chen, X., Francia, B., Li, M., 2004. Shared information and program plagiarism detection. IEEE. Trans. Inf. Theory 50 (7), 1545–1551.

Cilibrasi, R., Vitanyi, P., de Wolf, R., 2004. Algorithmic clustering of music based on string compression. Comput. Music J. 28 (4), 49–67.

Dai, Q., Yang, Y., Wang, T., 2008. Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison. Bioinformatics 24 (20), 2296–2302.

Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G., Fertil, B., 1999. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol. Biol. Evol. 16, 1391–1399.

Dufraigne, C., Fertil, B., Lespinats, S., Giron, A., Deschavanne, P., 2005. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acids Res. 33, e6.

Hamori, E., Ruskin, J., 1983. H curves, a novel method of representation of nucleotides series especially suited for long DNA sequences. J. Biol. Chem. 258, 1318–1327.

He, P., Wang, J., 2002. Characteristic sequences for DNA primary sequence. J. Chem. Inf. Comput. Sci. 42, 1080–1085.

Helden, J.V., 2004. Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics 20, 399–406.

Jeffrey, H.J., 1990. Chaos game representation of gene structure. Nucleic Acids Res. 18, 2163–2170.

Joseph, J., Sasikumar, R., 2006. Chaos game representation for comparison of whole genomes. BMC Bioinformatics 7, 243.

Kantorovitz, M., Robinson, G., Sinha, S., 2007. A statistical method for alignment free comparison of regulatory sequences. Bioinformatics 23, i249–i255.

Karlin, S., Burge, C., 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 11, 283–290.

Lake, J.A., 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl. Acad. Sci. USA 91, 1455–1459.

Li, C., Tang, N., Wang, J., 2006. Directed graphs of DNA sequences and their numerical characterization. J. Theor. Biol. 241, 173–177.

Li, C., Xing, L., Wang, X., 2008. 2-D graphical representation of protein sequences and its application to coronavirus phylogeny. BMB Rep. 41, 217–222.

Li, M., Badger, J., Chen, X., Kwong, S., Kearney, P., Zhang, H., 2001. An information based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics 17, 149–154.

Liao, B., 2005. A 2D graphical representation of DNA sequence. Chem. Phys. Lett. 401, 196–199.

Liu, N., Wang, T., 2005. A relative similarity measure for the similarity analysis of DNA sequences. Chem. Phys. Lett. 408, 307–311.

Liu, Y., Yang, Y., Wang, T., 2007. Characteristic distribution of L-tuple for DNA primary sequence. J. Biomol. Struct. Dyn. 25, 85–91.

Mantaci, S., Restivo, A., Rosone, G., Sciortino, M., 2007. An extension of the Burrows–Wheeler transform. Theor. Comput. Sci. 387, 298–312.

Mantaci, S., Restivo, A., Sciortino, M., 2008. Distance measures for biological sequences: some recent approaches. Int. J. Approx. Reason. 47, 1–18.

Nandy, A., 1994. A new graphical representation and analysis of DNA sequence structure. Curr. Sci. 66, 309–314.

Otu, H., Sayood, K., 2003. A new sequence distance measure for phylogenetic tree construction. Bioinformatics 19 (16), 2122–2130.

Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85, 2444–2448.

Pham, T., 2007. Spectral distortion measures for biological sequence comparisons and database searching. Pattern Recognition 40, 516–529.

Pham, T., Zuegg, J., 2004. A probabilistic measure for alignment-free sequence comparison. Bioinformatics 20, 3455–3461.

Pride, D.T., Meinersmann, R.J., Wassenaar, T.M., Blaser, M.J., 2003. Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Res. 13, 145–158.

Qi, J., Wang, B., Hao, B.L., 2004. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. J. Mol. Biol. 58, 1–11.

Qi, X., Wen, J., Qi, Z., 2007. New 3D graphical representation of DNA sequence based on dual nucleotides. J. Theor. Biol. 249, 681–690.

Randic, M., 2000. On the similarty of DNA primary sequences. J. Chem. Inf. Comput. Sci. 40, 50–56.

Randic, M., Guo, X., Basak, S.C., 2001. On the characterization of DNA primary sequences by triplet of nucleic acid bases. J. Chem. Inf. Comput. Sci. 41, 619–626.

Randic, M., Vracko, M., Lers, N., Plavsic, D., 2003a. Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem. Phys. Lett. 368, 1–6.

Randic, M., Vracko, M., Lers, N., Plavsic, D., 2003b. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. Chem. Phys. Lett. 371, 202–207.

Sandberg, R., Branden, C.I., Ernberg, I., Coster, J., 2003. Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G+C content. Gene 311, 35–42.

Sitnikova, T.L., Zharkikh, A.A., 1993. Statistical analysis of L-tuple frequencies in eubacteria and organells. BioSystems 30, 113–135.

Song, H., Tu, C., Zhang, G., et al., 2005. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. Proc. Natl. Acad. Sci. USA 102, 2430–2435.

Stuart, G.W., Moffett, K., Baker, S., 2002. Integrated gene and species phylogenies from unaligned whole genome protein sequences. Bioinformatics 18, 100–108.

Teeling, H., Meyerdierks, A., Bauer, M., Amann, R., Glockner, F.O., 2004. Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ. Microbiol. 6, 938–947.

Vinga, S., Almeida, J.S., 2003. Alignment-free sequence comparison—a review. Bioinformatics 19 (4), 513–523.

Wang, Y., Hill, K., Singh, S., Kari, L., 2005. The spectrum of genomic signatures: from dinucleotides to chaos game representation. Gene 346, 173–185.

Wu, T.J., Burke, J.P., Davison, D.B., 1997. A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. Biometrics 53, 1431–1439.

Wu, T.J., Hsieh, Y.C., Li, L.A., 2001. Statistical measures of DNA dissimilarity under Markov chain models of base composition. Biometrics 57, 441–448.

Yang, L., Zhang, X., Wang, T., 2010. The Burrows–Wheeler similarity distribution between biological sequences based on Burrows–Wheeler transform. J. Theor. Biol. 262, 742–749.

Yu, J., Sun, X., Wang, J., 2009. TN curve: a novel 3D graphical representation of DNA sequence based on trinucleotides and its applications. J. Theor. Biol. 261, 459–468.

Zhang, C., Zhang, R., Ou, H., 2003. The Z curve database: a graphic representation of genome sequences. Bioinformatics 19, 593–599.

Zheng, W., Chen, L., Qu, H., Gao, F., Zhang, C., 2005. Coronavirus phylogeny based on a geometric approach. Mol. Phylogenet. Evol. 36, 224–232.