MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **70** (2013) 417-430

ISSN 0340 - 6253

Analysis of DNA Sequences Based on the Fuzzy Integral

Shengli Zhang[a] , **Yusen Zhang**[b1] , **Ivan Gutman**[c]

[a]*Department of Mathematics, Xidian University Xi'an 710071, China*

[b]*School of Mathematics and Statistics, Shandong University at Weihai, Weihai 264209, China*

[c]*Faculty of Science, University of Kragujevac, P. O. Box 60, 34000 Kragujevac, Serbia*

(Received July 5, 2012)

Abstract

We introduce the analysis of DNA sequences based on the fuzzy integral, and compare it with some other existing methods. The similarity and phylogenetic analysis on two real data sets illustrate that the proposed approach is effective and feasible.

1 Introduction

Development of the nucleotide and protein sequencing technology have resulted in an explosive growth in the number of known DNA and protein sequences. It has raised many fundamental and challenging questions to modern biology. The elucidation of the evolutionary history of different species is a major concern to biological science. Early approaches to deal with it were mainly based on the alignment of a gene or protein sequence, but traditional alignment methods are computationally intensive and meaningless to whole genome comparison because each genome has its own genes and gene order. Ac-

[1]Corresponding author: zhangys@sdu.edu.cn

cordingly, there is an urgent need to develop new sequence analysis methods utilizing the ever-increasing genome data.

Some researchers explored many alignment–free methods for similarity and phylogenetic analysis. For instance, distance methods, maximal parsimony methods, maximum likelihood methods and Bayesian methods [1–11], each of which has its own range of applicability. Biologists and researchers are always trying to develop efficient methods for complex phylogenetic analysis. Zhang et al. [12] proposed to use gene content to measure the distance, which did not perform efficiently when the gene content of the organisms under study are very similar. Karlin et al. [13] proposed the dinucleotide relative abundance $\rho_{XY} = f_{XY}/f_X f_Y$ which discounts bias in G+C content and general base composition, where f_X denotes the frequency of nucleotide X, and f_{XY} denotes the frequency of dinucleotide XY. Information theory is also used for phylogenetic analysis [14–19]. Besides, some methods based on graphical representations of DNA sequences were put forward [20–29], which usually map a DNA sequence to a set of plots in 2D/3D space, and use some graphical invariants to characterize this sequence. These methods provide a simple way of viewing, sorting and comparing various gene structures. Motivated by their work, in this paper, we propose to take the fuzzy integral into account for analysis of DNA sequences.

The rest of this paper is organized as follows. We first discuss the feature vector of DNA sequences and some definitions of fuzzy measure and fuzzy integral, and then use the fuzzy integral similarity to obtain the distance metric. We finally apply the proposed method to two data sets: the coding sequences of the β-globin gene for 11 different species and the 24 coronavirus whole genomes. The similarity matrix and phylogenetic tree constructed by the new method are consistent with the commonly accepted ones. By comparing our method with other existing methods, we can see that these results are very promising and suggest more efforts for further developments.

2 Materials and methods

2.1 Feature vector of DNA sequences

Given a DNA sequence of length L, let $N(a_1a_2\ldots a_k)$ be the occurrences of a k-word $a_1a_2\ldots a_k$ observed in sequence, where a_i is one of the four nucleotides A, C, G or T and

k is the word length ($1 \leq k \leq L$). The frequency of $a_1a_2 \ldots a_k$ is defined by

$$f(a_1a_2 \ldots a_k) = N(a_1a_2 \ldots a_k)/(L-k+1)$$

Mutations happen in a more or less random manner at the molecular level, while selections shape the direction of evolution. From the perspective of molecular evolution, k-word frequency may reflect both the results of random mutation and selective evolution. One should reduce the random background from the simple counting result in order to highlight the contribution of selective evolution [13, 30, 31]. Here, we estimate the probability of random background by using the zeroth–order Markov model:

$$f^0(a_1a_2 \ldots a_k) = f(a_1)f(a_2) \cdots f(a_k)$$

where k ranges from 2 to L.

In this work, we collect

$$\alpha(a_1a_2 \ldots a_k) = \begin{cases} f(a_1a_2 \ldots a_k)/f^0(a_1a_2 \ldots a_k) & \text{if } f^0(a_1a_2 \ldots a_k) \neq 0 \\ 0 & \text{if } f^0(a_1a_2 \ldots a_k) = 0 \end{cases}$$

for all possible words $a_1a_2 \ldots a_k$ as the multi–nucleotide relative abundance of DNA sequence.

The selection of word length k is important to capture rich evolutionary information of DNA sequence. Weber et al. [32] and Reuben et al. [33] investigated the relationships among many important properties of genetic codon and 20 kinds of amino acids. They found that not only within the codon and amino acids, but also between codon and amino acids, there exist a number of significant correlations in nature. Meanwhile, codon-level phylogenetic analysis is the key topic in genome evolution, protein function and interactions between genetic and environment [34, 35]. Therefore, it will make sense to consider the importance of the triplet genetic code (k=3) in similarity and phylogenetic analysis of DNA sequences.

For a fixed $k = 3$, there are 64 distinct 3-words to be considered. By letting

$$\alpha_W = \sum_{\{X,Y\} \subseteq \{A,C,G,T\}} \alpha(XWY)$$

where $W \in \{A, C, G, T\}$, we get the *features vector* of DNA sequence, denoted as $(\alpha_A, \alpha_C, \alpha_G, \alpha_T)$.

For DNA sequences A and B, 4-word feature vectors $A = (\alpha_A^A, \alpha_C^A, \alpha_G^A, \alpha_T^A)$ and $B = (\alpha_A^B, \alpha_C^B, \alpha_G^B, \alpha_T^B)$ are constructed, that can be used to discriminate DNA sequences from different species.

2.2 Fuzzy measure and fuzzy integral

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a finite set, let $A, B \subseteq X$, and let $\Re(X)$ be the power set of X. A fuzzy measure, μ, is a real valued function $\mu : \Re(X) \rightarrow [0, 1]$, satisfying the following conditions:

(a) $\mu(\emptyset) = 0 \quad$ and $\quad \mu(X) = 1$

(b) $\mu(A) \leq \mu(B) \quad$ if $\quad A \subseteq B$.

The λ-fuzzy measure [36,37], that we use in this work, satisfies the properties of fuzzy measure plus the following additional condition: for all $A, B \subset X$ and $A \cap B = \emptyset$,

$$\mu(A \cup B) = \mu(A) + \mu(B) + \lambda\, \mu(A)\, \mu(B) \quad \text{for some } \lambda > -1 \tag{1}$$

where λ is obtained by solving the equation

$$\lambda + 1 = \prod_{i=1}^{n} (1 + \lambda\, \mu^i) \; . \tag{2}$$

Let $h : X \rightarrow [0, 1]$ represent a function that matches each element of X to its evidence. Suppose that $h(x_1) \geq h(x_2) \geq \cdots \geq h(x_n)$. If this is not the case for any element, then reorder X so that the relation holds, and let $\mu : \Re(X) \rightarrow [0, 1]$ be a fuzzy measure. Then the *fuzzy integral* of h with respect to the fuzzy measure μ is:

$$I = \max_{i=1}^{n} [\min(h(x_i), \mu(A_i))] \tag{3}$$

where $A_i = \{x_1, x_2, \ldots, x_i\}$.

2.3 Fuzzy integral similarity and distance metric

Let $A = (\alpha_A^A, \alpha_C^A, \alpha_G^A, \alpha_T^A)$ and $B = (\alpha_A^B, \alpha_C^B, \alpha_G^B, \alpha_T^B)$ be two normalized columns to be compared. Here, the so-called h function can be defined as $h(i) = 1 - |i^A - i^B|$, where $i = \{\alpha_A, \alpha_C, \alpha_G, \alpha_T\}$, i. e., the similarity of the feature vectors A and B.

Consider the maximum level of conservation of the feature vector, which favors the importance of better conserved positions. We can define a λ-fuzzy measure μ, in our case, $\mu^i = max(i^A, i^B)$. At this point, we can just apply Eq. (2) to obtain λ, and Eq. (1) to obtain the fuzzy measure μ. It can be easily proven that μ satisfies the conditions (a) and (b) of the fuzzy measures. Once we have h and μ, it is a straightforward task to obtain the fuzzy integral by using Eq. (3).

According to the fuzzy integral similarity measure, we can define the distance metric between two feature vectors. Given the feature vectors A and B, their distance is $D(A,B) = 1 - I(A,B)$.

It had been proved that the distance D satisfies the following four properties required by distance metrics:

(1) $D(A,B) > 0, \quad \forall A \neq B$;

(2) $D(A,B) = 0, \quad \forall A = B$;

(3) $D(A,B) = D(B,A), \quad \forall A, B$;

(4) $D(A,B) \leq D(A,C) + D(C,B), \quad \forall A, B, C$.

We will consider the feature vectors of DNA sequences and calculate their distances according to the above equation. By arranging all these values into a matrix, a pairwise distance matrix is derived. This distance matrix contains the similarity information on the n DNA primary sequences. Finally, this pairwise distance matrix may be input to the Neighbour program in PHYLIP package [38] for constructing a phylogenetic tree.

3 Experiments and Results

In order to test our method, we have selected two test data, the coding sequences of the β-globin gene for 11 different species and the 24 coronavirus whole genomes separately. The phylogenetic reconstruction of the two data sets using our new distance, all pointed at encouraging results.

3.1 Similarity analysis of the β-globin gene for 11 different species

In the first experiment, we choose the coding sequences of the β-globin gene for 11 different species, reported by Randić et al. [23]. Taxonomic information and accession numbers are provided in Table 1.

The similarity matrix M was obtained by the above specified and is shown in Table 2. It is based on the assumption that two DNA sequences are more similar if they have smaller least similarity values, which means that the corresponding least similarity value is close to 0.

Table 1. The accession numbers, length, and location for each β-globin genes and their exons

	Species	Database	ID	Location	Leng(bp)	Location of each exon		
1	Human	NCBI	U01317	62187-63610	1424	62187$\cdots$62278,	62409$\cdots$ 62631,	63482$\cdots$ 63610
2	Chimpanzee	NCBI	X02345	4189-5532	1344	4189$\cdots$4293,	4412$\cdots$4633,	5484$\cdots$5532
3	Gorilla	NCBI	X61109	4538-5881	1344	4538$\cdots$4630,	4761$\cdots$4982,	5833$\cdots$5881
4	Lemur	NCBI	M15734	154-1595	1442	154$\cdots$245,	376$\cdots$598,	1467$\cdots$1595
5	Rat	NCBI	X06701	310-1505	1196	310$\cdots$401,	517$\cdots$739,	1377$\cdots$1505
6	Mouse	NCBI	V00722	275-1462	1188	275$\cdots$367,	484$\cdots$705,	1334$\cdots$1462
7	Goat	NCBI	M15387	279-1749	1471	279$\cdots$364,	493$\cdots$715,	1621$\cdots$1749
8	Bovine	NCBI	X00376	278-1741	1464	278$\cdots$363,	492$\cdots$714,	1613$\cdots$1741
9	Rabbit	NCBI	V00882	277-1419	1143	277$\cdots$368,	495$\cdots$717,	1291$\cdots$1419
10	Opossum	NCBI	J03643	467-2488	2022	467$\cdots$558,	672$\cdots$894,	2360$\cdots$2488
11	Gallus	NCBI	V00409	465-1810	1346	465$\cdots$556,	649$\cdots$871,	1682$\cdots$1810

Table 2: The similarity matrix of the coding sequences of the β-globin gene of 11 species

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	0	0.01788	0.06387	0.12706	0.04632	0.02484	0.05827	0.03207	0.02446	0.02157	0.01208
Goat		0	0.06534	0.12559	0.03599	0.04239	0.05940	0.03320	0.02593	0.01386	0.01975
Opossum			0	0.19093	0.09124	0.06174	0.07406	0.06281	0.04164	0.07919	0.05179
Gallus				0	0.12804	0.12918	0.15145	0.14665	0.15152	0.11173	0.13913
Lemur					0	0.05267	0.03413	0.02944	0.04959	0.04068	0.04820
Mouse						0	0.03765	0.03025	0.03651	0.03690	0.02504
Rabbit							0	0.02619	0.04875	0.06409	0.04827
Rat								0	0.02256	0.03789	0.02207
Gorilla									0	0.03979	0.01238
Bovine										0	0.02740
Chimpanzee											0

From Table 2, we see that the smallest entries in it are associated with the pairs (Human, Chimpanzee), (Human, Gorilla), (Gorilla, Chimpanzee) and (Goat, Bovine). Furthermore, human is found to more similar to chimpanzee than gorilla. On the other hand, the largest entries in the similarity values appear in the rows belonging to opossum (the most remote species from the remaining mammals) and gallus (the only non-mammalian representative). This is consistent with the known facts of evolution.

In order to see this more clearly, in Fig. 1 we show the phylogenetic tree of the β-globin gene for 11 different species. Similar results have been obtained also elsewhere [20–24].

One should bare in mind that the values presented in Table 2 pertain to the comparison of multi-sequences, not to the comparison of sequences one by one. This means that the values in Table 2 only show the relative relations among these sequences, whereas the right phylogenetic relation among them should be established by additional algorithms. Different algorithms may result in different phylogenetic trees, so it is important to choose the most appropriate among them. In the present paper, the result in Fig 1. were generated by means of the UPGMA approach (UPGMA = **U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic Mean) [39].

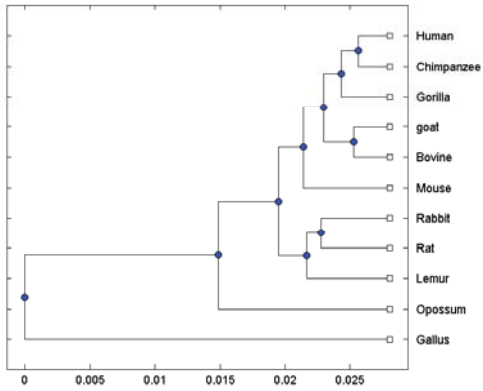

Fig. 1. The phylogenetic tree of the β-globin gene for 11 different species.

In order to compare our proposed method with other, recently reported representative methods, we examined the similarity degree between human and the other 10 species by five different approach, see Fig. 2. It is seen that our method is basically consistent with

the previous ones. Therefore we may conclude that the method proposed in this work is applicable for similarity and phylogenetic analysis of DNA sequences of different species.

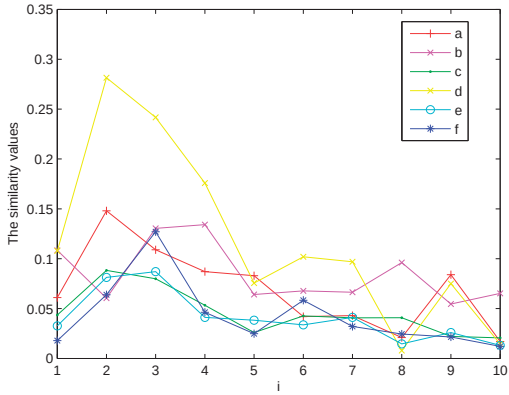

Fig. 2. The similarity degree comparison of the coding sequences of several species with the coding sequences of human (f: from Table 2 in the present work; a: from Randić et al. [23]; b: from Liao et al. [20]; c: from Liao et al. [21]; d: from Liu et al. [22]; e: from Wang et al. [24]). On the abscissa, i corresponds the $(i+1)$-th species in Table 2.

3.2 Phylogenetic analysis of the 24 coronavirus whole genomes

In order to further verify the validity of our method, we performed a phylogenetic analysis of sequences belonging to the 24 coronavirus whole genomes, which are listed in Table 3. Coronaviruses are members of a family of enveloped viruses that replicate in the cytoplasm of animal host cell. According to the type of the host, coronaviruses can be classified into three groups. Groups I and II contain mammalian viruses, whereas group III contains only avian viruses. After genome sequencing of some SARS-CoVs, much effort has been made to identify, by using molecular data, the phylogenetic position of SARS-CoVs in the coronavirus tree.

The phylogenetic tree for 24 coronavirus whole genomes was constructed by using the above described method, and is presented in Fig. 3. In order to compare our method with alignment method, we also construct the evolutionary tree by ClustalW method [40], which is a multiple sequence alignment program. The result is shown in Fig. 4.

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

No.	Accession	Abbreviation	Genome	Leng(bp)
1	NC_ 002645	HCoV_ 229E	Human coronavirus 229E	27317
2	NC_ 002306	TGEV	Transmissible gastroenteritis virus	28586
3	NC_ 003436	PEDV	Porcine epidemic diarrhea virus	28033
4	U00735	BCoVM	Bovine coronavirus strain Mebus	31032
5	AF391542	BCoVL	Bovine coronavirus isolate BCoV-LUN	31028
6	AF220295	BCoVQ	Bovine coronavirus strain Quebec	31100
7	NC_ 003045	BCoV	Bovine coronavirus	31028
8	AF208067	MHVM	Murine hepatitis virus strain ML-10	31233
9	AF201929	MHV2	Murine hepatitis virus strain 2	31276
10	AF208066	MHVP	Murine hepatitis virus strain Penn 97-1	31112
11	NC_ 001846	MHV	Murine hepatitis virus strain A59	31357
12	NC_ 001451	IBV	Avian infectious bronchitis virus	27608
13	AY278488	BJ01	SARS coonavirus BJ01	29725
14	AY278741	Urbani	SARS coronavirus Urbani	29727
15	AY278491	HKU-39849	SARS coronavirus HKU-39849	29742
16	AY278554	CUHK-W1	SARS coronavirus CUHK-W1	29736
17	AY282752	CUHK-Su10	SARS coronavirus CUHK-SulO	29736
18	AY283794	SIN2500	SARS coronavirus Sin2500	29711
19	AY283795	SIN2677	SARS coronavirus Sin2677	29705
20	AY283796	SIN2679	SARS coronavirus Sin2679	29711
21	AY283797	SIN2748	SARS coronavirus Sin2748	29706
22	AY283798	SIN2774	SARS coronavirus Sin2774	29711
23	AY291451	TW1	SARS coronavirus TW1	29729
24	NC_ 004718	TOR2	SARS coronavirus	29751

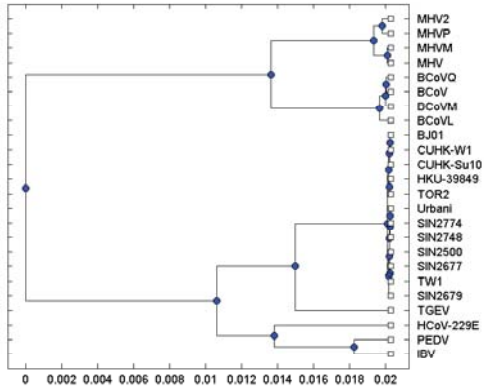

Fig. 3. The phylogenetic tree for 24 coronavirus whole genomes constructed by our method.

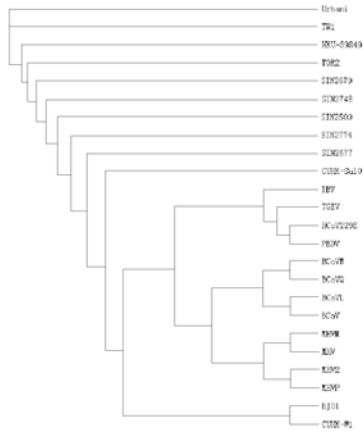

Fig. 4. The phylogenetic tree for 24 coronavirus whole genomes constructed by the ClustalW method.

Comparing the results shown in Figs. 3 and 4, we find that our method performs better: By our approach, coronaviruses are divided into four groups according to serotypes. Group I (HCoV 229E, TGEV, and PEDV) and group II (BCoVL, BCoVM, BCoVQ, BCoV, MHVM, MHV2, MHVP, and MHV) contain mammalian viruses, while group II coronaviruses contain a hemagglutinin esterase gene homologous to that of Influenza C

virus. Group III (IBV) contains only avian viruses, and Group IV [41, 42] are SARS-CoVs. From Fig. 3 we can observe that all the SARS-CoVs that belong to Group IV are clustered into the same class accurately. That is, all 12 SARS-CoV strains are grouped together and form a new fourth group, which is distinctly related to the group I coronaviruses (TGEV, explicitly). This is in accordance with the best result in the publicized existing trees [43]. An inspection of Fig.4 shows the grouping there is quite different. This corroborates the applicability of our method, relative to the ClustalW procedure.

4 Conclusions and Discussion

With the development of technology, more and more biological sequences are being collected for analysis. In the present study, we introduce a similarity and phylogenetic analysis of DNA sequences based on the fuzzy integral. The main advantage is that our approach can consider not only the similarity of feature vectors of DNA sequences, but also the relative importance of each occurrence within each feature vector. Furthermore, our method does not require any additional parameter. This makes it more robust and fully automated, thus avoiding the need to select parameters via expert knowledge or trial–and–error schemes. Experiments on the coding sequences of the β-globin gene for 11 different species, and for the 24 coronavirus whole genomes have both indicated that our proposed method is efficient and feasible.

In summary, in this paper we offer a novel method yielding reasonably good results in a rapid manner. Our method is not necessarily an improvement as compared to some existing ones, but rather an alternative. It does not require sequence alignment and the construction of tree models. Our tests have demonstrated that our method can serve as an alternative tool among other alignment–based and alignment–free approaches for similarity and phylogenetic analysis of DNA sequences.

Acknowledgement: This work was supported in part by Scientific Research Startup Foundation of Xidian University and the Fundamental Research Funds for the Central Universities, the Shandong Natural Science Foundation (Grant No. ZR2010AM020), and by the Serbian Ministry of Science and Education (Grant No. 174033).

References

[1] Y. Lin, S. Fang, J. Thorne, A tabu search algorithm for maximum parsimony phylogeny inference, *Eur. J. Oper. Res.* **176** (2007) 1908–1917.

[2] F. Ren, H. Tanaka, Z. Yang, A likelihood look at the supermatrix–supertree controversy, *Gene* **441** (2009) 119-125.

[3] A. Som, ML or NJ–MCL? A comparison between two robust phylogenetic methods, *Comput. Biol. Chem.* **33** (2009) 373–378.

[4] M. B. Elliott, D. M. Irwin, E. P. Diamandis, In silico identification and bayesian phylogenetic analysis of multiple new mammalian kallikrein gene families, *Genomics* **88** (2006) 591–599.

[5] E. Jako, E. Ari, P. Ittzes, A. Horvath, J. Podani, BOOL-AN: A method for comparative sequence analysis and phylogenetic reconstruction, *Mol. Phy. Evol.* **52** (2009) 887–897.

[6] Y. S. Zhang, W. Chen, A measure of DNA sequence dissimilarity based on free energy of nearest–neighbor interaction, *J. Biomol. Struct. Dyn.* **28** (2011) 557–565.

[7] X. Q. Qi, Q. Wu, Y. S. Zhang, E. Fuller, C. Q. Zhang, A novel model for DNA sequence similarity analysis based on graph theory, *Evol. Bioinformatics* **7** (2011) 149–158.

[8] Y. S. Zhang, W. Chen, A new measure for similarity searching in DNA sequences, *MATCH Commun. Math. Comput. Chem.* **65** (2011) 477–488.

[9] Y. Q. Liu, Y. S. Zhang, New invariant of DNA sequences based on a new matrix representation, *Comb. Chem. High. T. Scr.* **14** (2011) 61–71.

[10] H. L. Wang, Y. S. Zhang, A new approach to molecular phylogeny of H5N1 avian influenza viruses in Asia, *Int. J. Quantum Chem.* **110** (2010) 1964–1971.

[11] Y. Q. Liu, Y. S. Zhang, A new method for analyzing H5N1 avian influenza virus, *J. Math. Chem.* **47** (2010) 1129–1144.

[12] H. Zhang, Y. Zhong, B. Hao, X. Gu, A simple method for phylogenomic inference using the information of gene content of genomes, *Gene* **441** (2009) 163–168.

[13] S. Karlin, M. Ladunga, Comparisons of eukaryotic genomic sequences, *Proc. Natl. Acad. Sci.* **91** (1994) 12832–12836.

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

[15] D. R. Bastola, H. H. Otu, S. E. Doukas, K. Sayood, S. H. Hinrichs, P. C. Iwen, Utilization of the relative complexity measure to construct a phylogenetic tree for fungi, *Mycol. Res.* **108** (2004) 117–125.

[16] S. Zhang, L. Yang, T. Wang, Use of information discrepancy measure to compare protein secondary structures, *J. Mol. Struct. (THEOCHEM)* **909** (2009) 102–106.

[17] S. Zhang, T. Wang, Phylogenetic analysis of protein sequences based on conditional LZ complexity, *MATCH Commun. Math. Comput. Chem.* **63** (2010) 701–716.

[18] L. Yang, X. Zhang, T. Wang, The Burrows–Wheeler similarity distribution between biological sequences based on Burrows–Wheeler transform, *J. Theor. Biol.* **262** (2010) 724–749.

[19] S. Zhang, T. Wang, A complexity–based method to compare RNA secondary structures and its application, *J. Biomol. Struct. Dyn.* **28** (2010) 247–258.

[20] B. Liao, T. M. Wang, Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation, *Chem. Phys. Lett.* **388** (2004) 195–200.

[21] B. Liao, M. Tan, K. Ding, A 4D representation of DNA sequences and its application, *Chem. Phys. Lett.* **402** (2005) 380–383.

[22] X. Q. Liu, Q. Dai, Z. L. Xiu, T. M. Wang, PNN-curve: A new 2D graphical representation of DNA sequences and its application, *J. Theor. Biol.* **243** (2006) 555–561.

[23] 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.

[24] S. Y. Wang, F. C. Tian, W. J. Feng, X. Liu, Applications of representation method for DNA sequences based on symbolic dynamics, *J. Mol. Struct. (THEOCHEM)* **909** (2009) 33–42.

[25] Y. Guo, T. Wang, A new method to analyze the similarity of protein structure using TOPS representations, *J. Biomol. Struct. Dyn.* **26** (2008) 367–364.

[26] J. Feng, T. Wang, Condensed representations of protein secondary structure sequences and their application, *J. Biomol. Struct. Dyn.* **25** (2008) 621–628.

[27] Y. J. Huang, T. M. Wang, New graphical representation of a DNA sequence based on the ordered dinucleotides and its application to sequence analysis, *Int. J. Quantum Chem.* **112** (2012) 1746–1757.

[28] M. Randić, J. Zupan, A. T. Balaban, D. Vikić–Topić, D. Plavšić, Graphical representation of proteins, *Chem. Rev.* **111** (2011) 790–862.

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

[30] L. Gao, J. Qi, B. L. Hao, Simple Markov subtraction essentially improves prokaryote phylogeny, *Assoc Asia Pacific Phys. Soc. Bull.* **6** (2006) 3–7.

[31] J. Qi, B. Wang, B. L. Hao, Whole proteome prokaryote phylogeny without sequence alignment: A K-String composition approach, *J. Mol. Biol.* **58** (2004) 1–11.

[32] A. L. Weber, J. C. Lacey, Genetic code correlations: amino acids and their anticodon nucleotides, *J. Mol. Evol.* **11** (1978) 199–210.

[33] J. Reuben, F. Polk, Nucleotide–amino acid interactions and their relation to the genetic code, *J. Mol. Evol.* **15** (1980) 103–112.

[34] H. Goodarzi, N. Torabi, H. S. Najafabadi, M. Archetti, Amino acid and codon usage profiles: Adaptive changes in the frequency of amino acids and codons, *Gene* **407** (2008) 30–41.

[35] G. R. Moura, J. A. Paredes, M. A. S. Santos, Development of the genetic code: Insights from a fungal codon reassignment, *FEBS Lett.* **584** (2010) 334–341.

[36] M. Sugeno, Fuzzy measures and fuzzy integrals: A survey, in: M. M. Gupta, G. N. Saridis (Eds.), *Fuzzy Automata and Decision Processes*, North Holland, Amsterdam, 1977, pp. 89–102.

[37] F. Garcia, F. J. Lopez, C. Cano, A. Blanco, FISim: A new similarity measure between transcription factor binding sites based on the fuzzy integral, *BMC Bioinformatics* **10** (2009) #224.

[38] J. Felsenstein, PHYLIP-phylogeny inference package (version 3.2). *Cladistics* **5** (1989) 164–166.

[39] R. Sokal, C. Michener, A statistical method for evaluating systematic relationships, *Univ. Kansas Sci. Bull.* **38** (1958) 1409-1438.

[40] J. D. Thompson, D. G. Higgins, T. J. Gibson TJ, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice, *Nucleic Acids Res.* **22** (1994) 4673–4680.

[41] M. A. Marra et al. (58 coauthors), The genome sequence of the SARS-associated coronavirus, *Science* **300** (2003) 1399–1404.

[42] P. A. Rota et al. (33 coauthors), Characterization of a novel coronavirus associated with severe acute respiratory syndrome, *Science* **300** (2003) 1394–1399.

[43] J. Wang, X. Zheng, WSE, a new sequence distance measure based on word frequencies, *Math. Biosci.* **215** (2008) 78–83.