The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010
Copyright © 2010 ORSC & APORC, pp. 72–80

A Measure for Sequence Similarity Based on Dual Nucleotides and Information Discrepancy

Zhen Wang[1,*], Jia-Wei Luo[1], Han-Qi Zhou[1], Fang Liu[1]

1 School of Computer and Communication
Hunan University, Changsha, Hunan Province, 410082

Abstract The Function of Degree of Disagreement (FDOD) is a measurement of discrete information among multiple information sources, which has been successfully applied to phylogenetic analysis, structural classification of proteins, analysis of SARS virus, etc. According to the chemical property of the dual nucleotides, a measurement of sequence similarity based on the frequencies of the dual nucleotides is brought up. The phylogenetic relationships of the first exon of β-globin gene of 11 species and the 11 types of SARS coronavirus isolate genome sequence illustrate the utility of our approach.

Keywords analysis of similarity; FDOD; complete information set; dual nucleotides

1 Introduction

The main purpose of the sequence similarity analysis is to clarify the homology relationship between the sequences and forecast the structure and function of the new sequence from the known sequences. The complexity of multiple alignments is exponential with the growth of the number or length of the sequences. So for the whole-genome data, the alignment is hard to achieve[1]. Sequence alignment methods require the users to set parameters, penalty, insert the space to equal the sequence length. The some subjective factors lead to that different users may get the different results[2]. Response to the above inadequacies and limitations, some researchers advance a number of methods which can calculate the sequence similarity between different species without multiple alignments. With the development of information theory, information theory methods gained more attention. Professor Fang advanced a method of the Function of Degree of Disagreement (FDOD)[3][4]. The main character is using the sequence information to descript their relationship without any subjective information. FDOD has been successfully applied to phylogenetic analysis[5], structural classification of proteins[6][7], analysis of SARS virus[8], etc. However, in FDOD the complete information set of the sequence is represented by the probability distribution of l_subsequence ($l \geq 1$). The dimension of the l_subsequence increases exponentially with the increase of l. For a subsequence with the length is l, the dimension is 4^l（DNA sequence）, 20^l（protein sequence）. The

[*] Email: lance4@qq.com

dimension will be very large while l is larger. In this paper, we advance a new complete information set of sequence which can be representative of the information and features of DNA sequence. Complete information set of DNA sequence is composed of frequencies of the dual nucleotides in different intervals. For DNA sequence, there are only sixteen kinds of neighboring two bases, so the dimensions of the probability distributions of the neighboring two bases in different intervals are 16. This method can reduce the space complexity.

2 Similarity Analysis

2.1 FDOD

The Function of Degree of Disagreement (FDOD) is a measurement of discrete information based on information theory by Fang[9].

$\sum = \{a_1,a_2,...,a_m\}$ is an alphabet of m symbols. $S = \{S_1,S_2,...,S_S\}$ is a sequence set formed from the symbol set $\sum$. We denote the set of all different sequences formed from $\sum$ with length l by Θ^l , so the number $m(l)$ of all sequences of Θ^l equals m^l. For a sequence $S_k \in S$, Let the length of S_k be L_k and n^l_{ik} mean the number of the consecutive subsequence segments in S_k that match the ith sequence of Θ^l with length l. Obviously, the number of the subsequences with length l in S_k is:

$$\sum_{i=1}^{m(l)} n_{ik}{}^{l} = L_k - l + 1 \tag{1}$$

Letting $p_{ik}{}^{l} = \dfrac{n_{ik}{}^{l}}{L_k - l + 1}$, so we can obtain a distribution of a sequence $U_k{}^{l} = (p_{1k}{}^{l}, p_{2k}{}^{l}, ..., p_{m(l)k}{}^{l})^T$, where $\sum_{i=1}^{m(l)} p^l_{ik} = 1, 1 \le L_k$

For a given set of sequences S = {S1,S2,…,SS}, we can get the distribution of a set of subsequences:

$$\begin{aligned} &U_1{}^{l} = (p^l_{11}, p^l_{21}, ..., p^l_{m(l)1})^T \\ &U^l_2 = (p^l_{12}, p^l_{22}, ..., p^l_{m(l)2})^T \\ &... \\ &U^l_s = (p^l_{1s}, p^l_{2s}, ..., p^l_{m(l)s})^T \end{aligned} \tag{2}$$

Where $\sum_{i=1}^{m(l)} p^l_{ik} = 1,\ k = 1, 2, ..., S$

The FDOD function is defined as follows:

$$B(U_1{}^{l}, ..., U_S{}^{l}) = \sum_{k=1}^{S} \sum_{i=1}^{m(l)} p^l_{ik} \log \frac{p^l_{ik}}{\sum_{j=1}^{S} p^l_{ij} / S} \tag{3}$$

Where $0 \log 0=0$, $0 \log(0/0)=0$.

In a DNA primary sequence, $\sum = \{A, G, C, T\}$ and $m^l=4^l$. Suppose there are two DNA sequences X and Y, we can obtain discrepancies based on FDOD function:

$$\mathrm{D}(X,Y)=\sum_{j=X}^{Y}\sum_{i=1}^{m(l)} {p_{ij}}^{l}\log\frac{{p_{ij}}^{l}}{({p_{iX}}^{l}+{p_{iY}}^{l})/2} \tag{4}$$

D(X, Y) means discrepancies of two species. The larger D(X, Y) is, the lower the similarity of the two sequences is.

Using FDOD function to calculate discrepancies of DNA sequences, they only depend on the original information about sequences, no other subjective factors are involved[10].

2.2 A measure of sequence similarity based on dual nucleotides and information discrepancy

According to chemical properties of the adjacent nucleotide diad, we propose the measure of sequence similarity based on the frequency of dual nucleotides upon the FDOD function.

In a DNA primary sequence, the four DNA bases A, C, G and T can be divided into three classes by their chemical properties: purine R = {A, G}/pyrimidine Y = {C,T}, amino M = {A, C}/ketone K = {G, T}, and weak hydrogen bond W = {A, T}/strong hydrogen bond S = {C, G}. By considering neighboring two bases and the base order, we can get sixteen combinations: AG, GA, CT, TC, AC, CA, GT, TG, AT, TA, CG, GC, AA, CC, GG and TT.

According to the three classifications of the four DNA bases, the dual nucleotides can be divided into four classes: purine dual nucleotides {AG, GA}/pyrimidine dual nucleotides {CT, TC}, amino dual nucleotides {AC, CA}/ketone dual nucleotides {TG, GT}, weak hydrogen bond dual nucleotides {AT, TA}/strong hydrogen bond dual nucleotides {CG, GC}, and repeat dual nucleotides {AA, CC, GG, TT}. In each class, the frequencies of dual nucleotides in l interval are defined as following:

$$\begin{aligned}
&{dn^{l}}_{1}=AG\%=\frac{{AG^{l}}_{n-1}}{{AG^{l}}_{n-1}+{GA^{l}}_{n-1}+{CT^{l}}_{n-1}+{TC^{l}}_{n-1}} &&{dn^{l}}_{9}=AT\%=\frac{{AT^{l}}_{n-1}}{{AT^{l}}_{n-1}+{TA^{l}}_{n-1}+{GC^{l}}_{n-\ 1}+{CG^{l}}_{n-\ 1}}\\
&{dn^{l}}_{2}=GA\%=\frac{{GA^{l}}_{n-1}}{{AG^{l}}_{n-1}+{GA^{l}}_{n-1}+{CT^{l}}_{n-1}+{TC^{l}}_{n-1}} &&{dn^{l}}_{10}=TA\%=\frac{{TA^{l}}_{n-1}}{{AT^{l}}_{n-1}+{TA^{l}}_{n-1}+{GC^{l}}_{n-1}+{CG^{l}}_{n-1}}\\
&{dn^{l}}_{3}=CT\%=\frac{{CT^{l}}_{n-1}}{{AG^{l}}_{n-1}+{GA^{l}}_{n-1}+{CT^{l}}_{n-1}+{TC^{l}}_{n-1}} &&{dn^{l}}_{11}=GC\%=\frac{{GC^{l}}_{n-1}}{{AT^{l}}_{n-1}+{TA^{l}}_{n-1}+{GC^{l}}_{n-1}+{CG^{l}}_{n-1}}\\
&{dn^{l}}_{4}=TC\%=\frac{{TC^{l}}_{n-1}}{{AG^{l}}_{n-1}+{GA^{l}}_{n-1}+{CT^{l}}_{n-1}+{TC^{l}}_{n-1}} &&{dn^{l}}_{12}=CG\%=\frac{{CG^{l}}_{n-1}}{{AT^{l}}_{n-1}+{TA^{l}}_{n-1}+{GC^{l}}_{n-1}+{CG^{l}}_{n-1}}\\
&{dn^{l}}_{5}=AC\%=\frac{{AC^{l}}_{n-1}}{{AC^{l}}_{n-1}+{CA^{l}}_{n-1}+{GT^{l}}_{n-1}+{TG^{l}}_{n-1}} &&{dn^{l}}_{13}=AA\%=\frac{{AA^{l}}_{n-1}}{{AA^{l}}_{n-1}+{GG^{l}}_{n-1}+{CC^{l}}_{n-1}+{TT^{l}}_{n-1}}\\
&{dn^{l}}_{6}=CA\%=\frac{{CA^{l}}_{n-1}}{{AC^{l}}_{n-1}+{CA^{l}}_{n-1}+{GT^{l}}_{n-1}+{TG^{l}}_{n-1}} &&{dn^{l}}_{14}=GG\%=\frac{{GG^{l}}_{n-1}}{{AA^{l}}_{n-1}+{GG^{l}}_{n-1}+{CC^{l}}_{n-1}+{TT^{l}}_{n-1}}\\
&{dn^{l}}_{7}=GT\%=\frac{{GT^{l}}_{n-1}}{{AC^{l}}_{n-1}+{CA^{l}}_{n-1}+{GT^{l}}_{n-1}+{TG^{l}}_{n-1}} &&{dn^{l}}_{15}=CC\%=\frac{{CC^{l}}_{n-1}}{{AA^{l}}_{n-1}+{GG^{l}}_{n-1}+{CC^{l}}_{n-1}+{TT^{l}}_{n-1}}\\
&{dn^{l}}_{8}=TG\%=\frac{{TG^{l}}_{n-1}}{{AC^{l}}_{n-1}+{CA^{l}}_{n-1}+{GT^{l}}_{n-1}+{TG^{l}}_{n-1}} &&{dn^{l}}_{16}=TT\%=\frac{{TT^{l}}_{n-1}}{{AA^{l}}_{n-1}+{GG^{l}}_{n-1}+{CC^{l}}_{n-1}+{TT^{l}}_{n-1}}
\end{aligned} \tag{5}$$

where ${AG^{l}}_{n-1}$, ${GA^{l}}_{n-1}$, ${CT^{l}}_{n-1}$, ${TC^{l}}_{n-1}$... and ${TT^{l}}_{n-1}$ respectively are the cumulative occurrence numbers of AG, GA, CT, TC,... and TT from the first to the (n-1)-*th* base in DNA sequence while the interval is l. The length of the sequence is n.

For the sixteen kinds of order dual base pairs in DNA sequence, the probability distribution dimensions of different intervals are the same. The novel complete information set of the DNA sequence can be defined by:

$$dn^L = \left(\begin{array}{c|cccc} & l=1 & l=2 & \cdots & l=L-1 \\ \hline dn_1 & dn_1^1 & dn_1^2 & \cdots & dn_1^{L-1} \\ dn_2 & dn_2^1 & dn_2^2 & \cdots & dn_2^{L-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ dn_{16} & dn_{16}^1 & dn_{16}^2 & \cdots & dn_{16}^{L-1} \end{array} \right) \tag{6}$$

where the length of the DNA sequence is *L***.**

Suppose that there are two DNA sequences X and Y, the corresponding information sets are dn_X^l and dn_Y^l with length *l*.

$$dn_X{}^l = \begin{pmatrix} dn_{1X}^1 & dn_{1X}^2 & \cdots & dn_{1X}^l \\ dn_{2X}^1 & dn_{2X}^2 & \cdots & dn_{2X}^l \\ \vdots & \vdots & \ddots & \vdots \\ dn_{16X}^1 & dn_{16X}^2 & \cdots & dn_{16X}^l \end{pmatrix} \qquad dn_Y{}^l = \begin{pmatrix} dn_{1Y}^1 & dn_{1Y}^2 & \cdots & dn_{1Y}^l \\ dn_{2Y}^1 & dn_{2Y}^2 & \cdots & dn_{2Y}^l \\ \vdots & \vdots & \ddots & \vdots \\ dn_{16Y}^1 & dn_{16Y}^2 & \cdots & dn_{16Y}^l \end{pmatrix} \tag{7}$$

According to FDOD the information divergence of dn_X^l and dn_Y^l is defined by:

$$D(dn_X{}^l, dn_Y{}^l) = \sum_{j=X}^{Y} \sum_{i=1}^{16} \sum_{k=1}^{l} dn_{ij}^k \log \frac{dn_{ij}^k}{(\sum_{t=X}^{Y} dn_{it}^k)/2} \tag{8}$$

Where 0 log0=0, 0 log(0/0)=0.

Obviously, the smaller D (dn_X^l, dn_Y^l) is, the more similar the DNA sequences are.

The complete information set is expressed by the probability distributions of *l*-subsequence in FDOD. The dimension of *l*-subsequence is the exponential growth. For DNA sequence the dimension is 4^l. In this paper, we propose the method of sequence similarity based on dual nucleotides and information dispersion. The space complexity is O(16**l*). So, the space complexity becomes lower by this measure on the work of sequence similarity.

3 Data sets and Results

3.1 Data sets

In this paper, we will use two sets of data as our test data. The first set is the first exon of β-globin gene of 11 species, which are listed in Table 1. These sequences are very conservative and short. As the classic testing data sample, it have been used in the comparison and analysis of DNA sequence in many literatures. The second set is the whole genome shotgun sequences of 11 SARS corona virus isolates, whose length are about 29700bp respectively. The detailed information is shown in Table 2.

Table 1 The coding sequences of the first exon of β-globin gene of 11 different species

Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG

Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG

Table 2 The species, accessions, lengths and origins of the 11 types of SARS coronavirus isolate genome sequence

No.	Species	Accession	Length(bp)	From
1	CUHK-Su10	AY282752	29736	Hong Kong
2	HKU-39849	AY278491	29742	Hong Kong
3	CUHK-W1	AY278554	29736	Hong Kong
4	Sin2500	AY283794	29711	Singapore
5	Sin2677	AY283795	29705	Singapore
6	Sin2679	AY283796	29711	Singapore
7	Sin2748	AY283797	29706	Singapore
8	Sin2774	AY283798	29711	Singapore
9	TOR2	AY274119	29751	Canada Toronto
10	TW1	AY291451	29729	Taiwan
11	Urbani	AY278741	29727	American

3.2 Results and Discussion

We use the measure of sequence similarity based on doubling nucleotides and information dispersion to build the information divergence matrix for the first exon of β-globin gene of 11 different species in Table 3 (L=10). Based on the information divergence matrix, using the Neighbor program in the PHYLIP package (http://evolution.genetics.washington.edu/phylip.html), we can obtain the phylogenetic tree belonging to eleven different species listed in Fig. 1. We build the information divergence matrix for the SARS coronavirus data in Table 4 (L=100) and obtain the phylogenetic tree in Fig. 2.

Table 3 Similarity/Dissimilarity Matrix for the Coding Sequences of Table 1 based on the information divergence (L=10)

Name.	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chim.
Human	0.0000	0.1294	0.1789	0.2401	0.1673	0.1387	0.1266	0.1523	0.0102	0.1186	0.0366
Goat		0.0000	0.2662	0.2473	0.1837	0.1575	0.1470	0.1825	0.1162	0.0342	0.1242
Opossum			0.0000	0.2403	0.2605	0.2104	0.2853	0.2196	0.2009	0.2579	0.1970
Gallus				0.0000	0.3653	0.3103	0.3360	0.3018	0.2326	0.2925	0.2756
Lemur					0.0000	0.1746	0.1221	0.1909	0.1748	0.1405	0.1615
Mouse						0.0000	0.1898	0.1392	0.1334	0.1546	0.1256
Rabbit							0.0000	0.2342	0.1210	0.1188	0.1226
Rat								0.0000	0.1596	0.1728	0.1533
Gorilla									0.0000	0.1096	0.0324
Bovine										0.0000	0.1108
Chim.											0.0000

Table 4 Similarity/Dissimilarity Matrix for the Coding Sequences of Table 2 (L=100)

No.	1	2	3	4	5	6	7	8	9	10	11
1	0.00000	0.00014	0.00008	0.00023	0.00042	0.00023	0.00033	0.00025	0.00005	0.00026	0.00032
2		0.00000	0.00019	0.00024	0.00047	0.00026	0.00034	0.00028	0.00011	0.00023	0.00028
3			0.00000	0.00030	0.00048	0.00031	0.00039	0.00034	0.00012	0.00033	0.00038
4				0.00000	0.00012	0.00003	0.00010	0.00003	0.00027	0.00005	0.00007
5					0.00000	0.00013	0.00016	0.00014	0.00047	0.00014	0.00016
6						0.00000	0.00011	0.00004	0.00027	0.00005	0.00008
7							0.00000	0.00011	0.00036	0.00014	0.00016
8								0.00000	0.00028	0.00007	0.00008
9									0.00000	0.00025	0.00031
10										0.00000	0.00004
11											0.00000

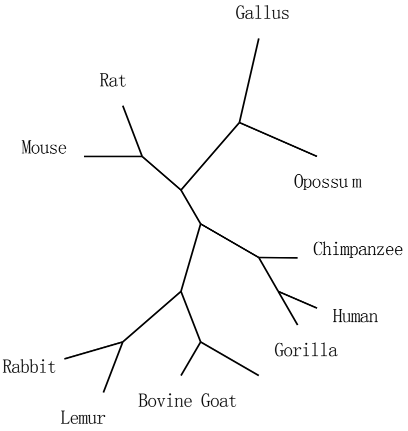

Fig. 1 Phylogenetic tree constructed by neighbor joining method

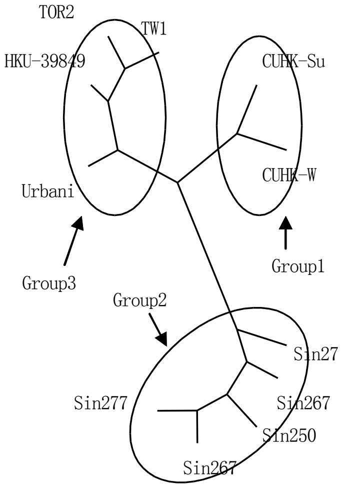

Fig. 2 Phylogenetic tree constructed by neighbor joining method for SARS coronavirus

From Table 3, we can see Gallus is farthest away from the other 10 species (the distances were all more than 0.2326, and the highest value reached 0.3653), followed by Opossum. Gallus is non-mammals while other species are mammals and Opossum has the furthest genetic relationship of all mammals. The similarity results are basically consistent with the facts. We find that the smaller entries in Table 2 are associated with the pairs Human and Gorilla(0.0102), Gorilla and Chimpanzee(0.0324), Goat and Bovine(0.0342), Human and Chimpanzee(0.0366). The discrepancy indices of these pairs are the smallest ones in the table, which means they are similar to each other in each pairs. It is more in line with the facts.

Then we give the analysis of the genetic relationship for the 11 SARS corona virus isolates. From Fig.2 we can see that group 1 (CUHK-Su10 and CUHK-W1) has a close genetic relationship, group 2 (Sin2748, Sin2677, Sin2500, Sin2679 and

Sin2774) has close affinity, group 3 (TOR2, TW1, HKU-39849 and Urbani) has close affinity from Fig.2. Referring to Table 2 it is distinct the types of group 1 come from Hong Kong, with the types of group 2 come from Singapore, while the group 3 can be attributed to global broadcasters. So, the genetic relationship for the 11 SARS corona virus isolates can be proved by their geography relationship.

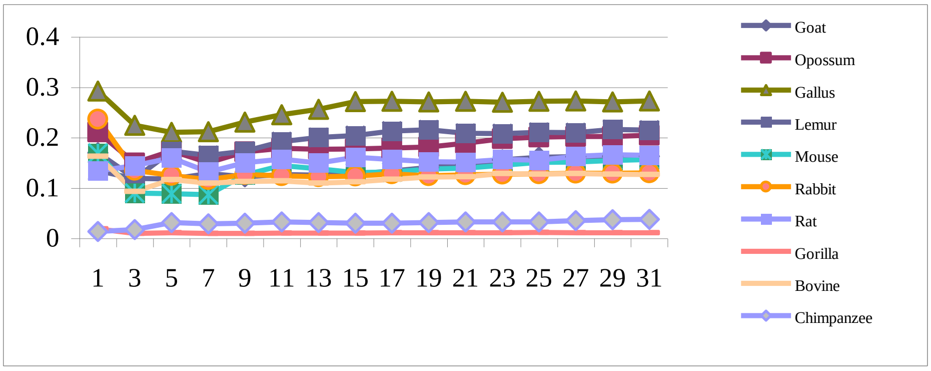

Fig. 3 The track changes of the distance values of Human and other 10 kinds of sequences with *L*

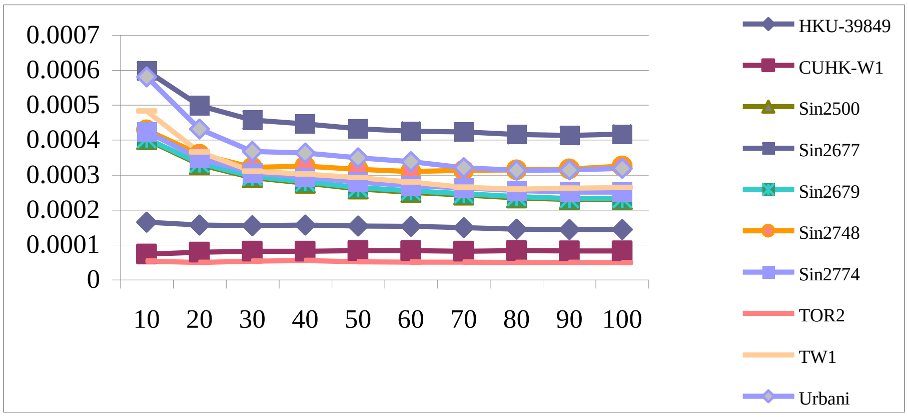

Fig. 4 The track changes of the distance values of CUHK-Su10 and other 10 kinds of viruses with *L*

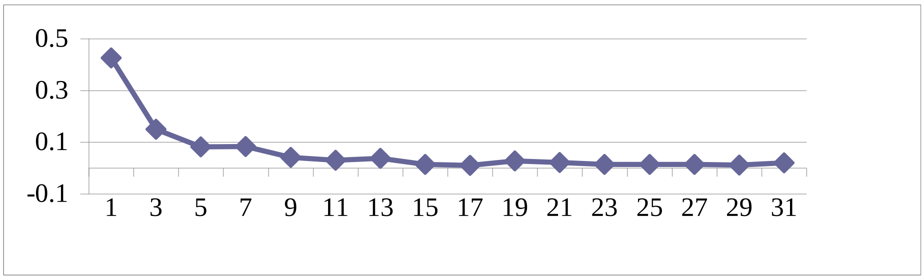

Fig. 5 The track changes of *bodong(L)* of Human with *L*

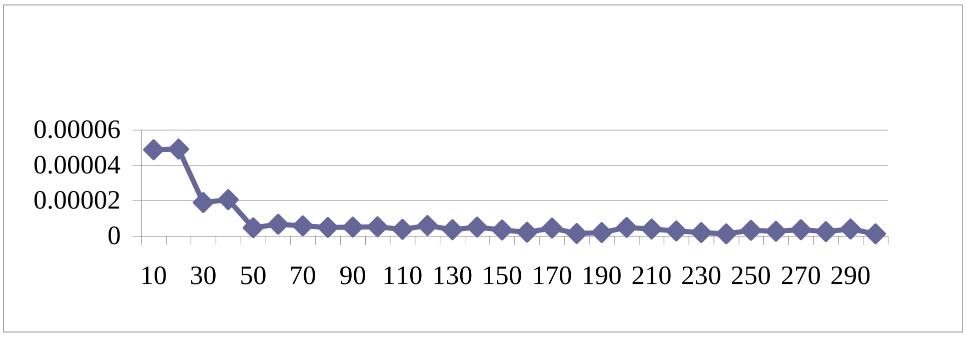

Fig. 6 The track changes of *bodong(L)* of CUHK-Su10 with *L*

Now we will study the relationship of the parameter *L* with the distance values between DNA sequences. Given an information set *P*, the distance matrix based on *P* is denoted by *dis*(*L*). When P^l changes to P^{l+1}, we define *bodong*(*L*) as the sum of the absolute value of difference between the corresponding elements of two distance matrices, i.e.,

$$bodong\,(L) = \sum |\, dis\,(L+1) - dis\,(L)\,| \tag{9}$$

Fig 3 shows that distance values of Human and other 10 kinds of sequences are changed with *L* while Fig 5 shows that the values of *bodong(L)* of Human are changed with *L*. Fig 4 shows that distance values of CUHK-Su10 and other 10 kinds of viruses are changed with *L* while Fig 6 shows that the values of *bodong(L)* of CUHK-Su10 are changed with *L*. From Fig 3 and Fig 5, we can see the distance values between species are not monotonically increasing or decreasing as the value of *L* increases. While *L* is very small, the information set include less information, so the distance values changed greatly. They tend to slowly change when *L* increases. Fig 4 and Fig 6 also proved this viewpoint. The results shown from Fig 3 to Fig 5 also indicate that we can set the proper value for parameter *L* only considering the variation of the distance value between DNA sequences without analyzing the complete information set *P* which could contend more information as the *L* increases.

4 Conclusion

In this paper, we consider the properties of the neighboring dual nucleotides to construct the novel complete information set of the DNA sequence based on the frequencies of the dual nucleotides in different intervals. According to information dispersion, we get an approach to make similarity analysis of DNA sequence. The advantage of our method is that can reflect the relationship between biological sequences objectively and computational complexity can be linear growth with the growth of sequence length. So, it is easy to process large scale data.

Acknowledgment

This work is supported in part by the National Nature Science Foundation of Hunan province (Grant No. 07JJ5086) and Hunan University Students innovative pilot scheme.

References

[1] Xu Cai, Weiwu Fang, Wen Zhang. Comparison of alignment-free methods based on mitochondrion complete genome. Computers and Applied Chemistry, 2005, 22(10): 837-844.

[2] Wen Zhang, Huanwen Tang, Weiwu Fang, Zhilong Xiu. Application of a new measure of information discrepancy to the analysis of SARS corona virus. Computers and Applied Chemistry, 2003, 20(6):719-732.

[3] Wei-Wu Fang, FS Roberts, Zhengrong Ma. A measure of discrepancy of multiple sequences. Information Science, 2001, 137:75-102.

[4] Wei-Wu Fang. The characterization of a measure of information discrepancy. Information Science, 2000, 125: 207-232.

[5] Wen Zhang, Huanwen Tang, Weiwu FANG, Xu Cai, Weiwei Zhang. Construction of phylogenetic tree of whole proteome microbial organisms as inferred from a new measure of information discrepancy. Journal of Dalian University of Technology, 2005, 45(6): 925-930.

[6] Li-Zhen Zhang, Huanwen Tang. A method of protein structure class prediction based on subsequence distribution. Computers and Applied Chemistry, 2003, 23(3):1-6.

[7] Jie Song, Huanwen Tang. Classification of Homo-oligomeric Proteins by a New Measure of Information Discrepancy. Mathematics in Practice and Theory, 2007, 37(8): 36-42.

[8] Zhen Qi, Runsheng Chen. Phylogenetic analysis of SARS coronavirus based on whole-genome comparison. Chinese Science Bulletin, 2003, 48(12): 1242-1245.

[9] WW Fang. The disagreement degree of multi-person judgments in additive structure. Mathematical Social Sciences, 1994, 28(2): 85-111.

[10] Min Zhang. Research on Multiple sequence Alignment Algorithms in Bioinformatics. Dalian University of Technology. 2005

[11] Ruan Y J, Wei C L, Ee L A, et al. Comparative full-length genome sequence analysis of 14 SARS coronavirus Isolates and common mutations associated with putative origins of infection. The Lancet, 2003, 361 (9371):1779-1785