J Math Chem (2011) 49:1479–1492
DOI 10.1007/s10910-011-9811-x

ORIGINAL PAPER

Phylogenetic analysis of DNA sequences with a novel characteristic vector

Yujuan Huang · Tianming Wang

Received: 11 January 2010 / Accepted: 19 March 2011 / Published online: 31 March 2011

Abstract In the basic biological research, one of major tasks is to compare biological sequences to infer evolutionary relations among sequences. In this paper, considering both the positions and numbers of a k-word and the random background, a novel characteristic vector of a DNA sequence is proposed to serve for genetic sequences comparison and phylogenetic analysis. The vector is composed of elements which characterize the relative difference of a DNA sequence from a sequence generated by a $(k-2)$th order Markov process. Finally, we reconstruct the phylogenetic trees of 48 HEV (Hepatitis E virus) and 20 Eutherian mammals. The results show that this new method provides more information about k-word and improves the efficiency of sequence comparison.

Keywords Biological sequence · Sequence comparison · Probability distribution · Phylogenetic tree

1 Introduction

The rapid development of biological information and technologies led to a great accumulation of DNA primary sequence database. Biological information obtained from

Y. Huang (✉) · T. Wang
School of Mathematical Sciences, Dalian University of Technology,
Dalian 116024, People's Republic of China
e-mail: yujuanh518@163.com

T. Wang
e-mail: wangtm@dlut.edu.cn

Y. Huang
Department of Mathematics & Physics, Shandong Jiaotong University,
Jinan 250023, People's Republic of China

biological sequences can prompt the development of genetic engineering and pharmaceutical research. However, it is very difficult to obtain directly information from DNA strands of different species. Therefore, in order to timely provide very useful information and insights for basic research and drug design, many kinds of methods surge to analyze and characterize a DNA sequence, among which methods of primary sequence comparison are useful assessment tools and play an important role in biological sequence analysis.

Early approaches on sequence comparison were mainly based on the direct alignment between gene sequences. So far, sequence alignment is still a very significant and useful tool in comparison for DNA primary sequence. Waterman [1] and Durbin et al. [2] provided comprehensive reviews about this method. The main idea of alignment was using a distance function to represent insertion, deletion, and substitution of letters in the compared sequences. However, the algorithm of sequence alignment encountered a lot of difficulties in computational aspect with respect to large biological databases and long sequences. Therefore, it is necessary to develop alignment-free methodologies and technologies to overcome critical limitations of sequence analysis by alignment.

In recent years, an increasing number of alignment-free methods have been proposed. As a very powerful tool for the visualization and analysis of DNA sequences, graphical representations are very useful [3–11]. These methods provide a simple way of viewing, sorting and comparing various gene structures. Besides, there have been a lot of computational and statistical methods for sequence comparisons. Comparison methods based on k-word frequencies [12–15] may be the most well-developed alignment-free ones. These methods utilized some concepts of distance measure [12], such as the Euclidean distance [16], Mahalanobis distances [17], and Kullback-Leibler discrepancy (KLD) [18] , Cosine distance [19] and so on, to serve for construction of phylogenetic trees. Another efficient statistical approach based on Markov model to whole genome comparison and phylogenetic analysis is the string composition vector (CV) proposed by Hao's group [20,22]. This method was initially used to infer prokaryote phylogeny and later also used to unicelluar eukaryotes such as fungi [23]. In this method, considering that mutations have been taking place randomly at molecular level and natural selections shape the direction of evolution, the composition vector was generated by subtracting the random background from the simple counting result to highlight the contribution of selective evolution [20].

Most methodologies with regard to word frequency are based on the numbers of the occurrence of a k-word and ignore its position in a DNA sequence. However, the positions of a k-word is closely connected with gene rearrangement, inversion, transposition, and translocation at the genome level. In this work, regarding a DNA sequence as a Markov process, we define a novel characteristic representation of a DNA sequence based on a new definition of probability distribution of a k-word. For the new definition, we simultaneously focus on the positions and numbers of a k-word. Motivated by the idea of CV, we derive a characteristic representation on differences of each k-word by subtracting the random background. Finally, we obtain a 4^k-dimensional vector, which can characterize the sequence, by collecting the maximum of the absolute value of each k-word's characteristic representation. The final distance for phylogenetic analysis between species is evaluated based on the cosine function between their corresponding composition vectors. As to its application, we reconstruct the phylogenetic trees of 48 HEV (Hepatitis E virus) and 20 Eutherian mammals.

2 Materials and methods

2.1 Probability distribution of a k-word

In many biological studies, a DNA sequence is interpreted as a succession of symbols from a finite alphabet $\mathscr{A} = \{A, C, G, T\}$. Let $S = s_1 s_2 \cdots s_L$ be a DNA sequence of length L. A k-word $w = \alpha_1\alpha_2 \cdots \alpha_k$ is a subsequence of k adjacent letters, $\alpha_i \in \mathscr{A}$. Obviously, there are a total of 4^k possible k-words for the alphabet $\mathscr{A}$. Let W_k denotes the set of all possible k-words, i.e.

$$W_k = \{w_1, w_2, \ldots, w_{4^k}\}$$

For a k-word w_j, p_i denotes the position of the ith occurrence of the w_j in S. We take into account both the position and number of a k-word and give the following new definition of $\tilde{P}_j^i(\alpha_1\alpha_2 \cdots \alpha_k)$ which can measure the probability of a k-word w_j occurring at the position p_i of S:

$$\tilde{P}_j^i(\alpha_1\alpha_2 \cdots \alpha_k) = \frac{p_i - p_{i-1}}{p_i + k - 1} \cdot \frac{i}{L - k + 1}, \quad 1 \le i \le m. \tag{1}$$

Here, m denotes the number of the occurrence of the k-word w_j and $p_0 = 0$. i denotes the number of the jth k-word until the position p_i. Note that $\tilde{P}_j^i$ is actually not a probability distribution since the sum of $\tilde{P}_j^i$ for all i is not equal to 1. We normalize it here by dividing the sum: $P_j^i(\alpha_1\alpha_2 \cdots \alpha_k) = \frac{\tilde{P}_j^i}{\sum_i \tilde{P}_j^i}$. So we derive a probability distribution P_j^i for each k-word.

2.2 Subtraction of random background and evolutionary distance measure

From the perspective of molecular evolution, the collection of the probability P_j^i may reflect both the result of random mutations and selective evolution. Mutations have been taking place randomly at molecular level and natural selections shape the direction of evolution. Many neutral mutations may remain and play a role of random background [21]. In order to highlight the selective diversification of sequence composition, we must subtract a random background which is done by using a $(k-2)$-th order Markovian prediction from the probability $P_j^i(\alpha_1\alpha_2 \cdots \alpha_k)$ for all i. The background probability of the jth k-word is defined as follows:

$$\bar{P}_j^i(\alpha_1\alpha_2 \cdots \alpha_k) = \frac{P_1^{k-1}(\alpha_1\alpha_2 \cdots \alpha_{k-1}) P_2^k(\alpha_2\alpha_3 \cdots \alpha_k)}{P_2^{k-1}(\alpha_2\alpha_3 \cdots \alpha_{k-1})}, \quad i = 1, 2, \ldots, m. \tag{2}$$

Here, $P_1^{k-1}(\alpha_1\alpha_2\cdots\alpha_{k-1})$, $P_2^k(\alpha_2\alpha_3\cdots\alpha_k)$ and $P_2^{k-1}(\alpha_2\alpha_3\cdots\alpha_{k-1})$ indicate the probability of the word $\alpha_1\alpha_2\cdots\alpha_{k-1}, \alpha_2\alpha_3\cdots\alpha_k$ and $\alpha_2\alpha_3\cdots\alpha_{k-1}$ respectively occurring at the position p_i, p_i+1 and p_i+1. Their definitions are similar to the definition of $\tilde{P}_j^i(\alpha_1\alpha_2\cdots\alpha_k)$. For example, $P_1^{k-1}(\alpha_1\alpha_2\cdots\alpha_{k-1}) = \frac{p_i-p_{i-1}^{'}}{p_i+(k-1)-1}\cdot\frac{i^{'}}{L-(k-1)+1}$. Here, $p_{i-1}^{'}$ denotes the nearest position of the word $\alpha_1\alpha_2\cdots\alpha_{k-1}$ to the position p_i. $i^{'}$ denotes the number of the occurrence of the word $\alpha_1\alpha_2\cdots\alpha_{k-1}$ until the position p_i. In a similar way, we can define the two other probability i.e. $P_2^k(\alpha_2\alpha_3\cdots\alpha_k)$ and $P_2^{k-1}(\alpha_2\alpha_3\cdots\alpha_{k-1})$. Thus, the difference between the probability P_j^i and the background probability $\bar{P}_j^i$ of the jth k-word at the position p_i is derived from the following definition:

$$\bar{\delta}_j^i(\alpha_1\alpha_2\cdots\alpha_k) = \begin{cases} \frac{P_j^i(\alpha_1\alpha_2\cdots\alpha_k)-\bar{P}_j^i(\alpha_1\alpha_2\cdots\alpha_k)}{\bar{P}_j^i(\alpha_1\alpha_2\cdots\alpha_k)}, & if\ \bar{P}_j^i \neq 0, \\ 0, & if\ \bar{P}_j^i = 0. \end{cases} \tag{3}$$

$i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, 4^k$.

Without collecting all the value of $\bar{\delta}_j^i$ for all i and j, we normalize $\bar{\delta}_j^i, i = 1, 2, \ldots, m$, by dividing their sum and derive the new value $\delta_j^i, i = 1, 2, \ldots, m$. $\{\delta_j^1, \delta_j^2, \cdots, \delta_j^m\}$ is called the characteristic representation of a k-word. Then we choose the maximum of the absolute value of $\{\delta_j^1, \delta_j^2, \ldots, \delta_j^m\}$, denoted by m_j, and collect all these maximums for each k-word to develop a 4^k-dimensional vector to characterize a DNA sequence.

For two species A and B, we form two vectors $A = (m_1, m_2, \ldots, m_{4^k})$ and $B = (n_1, n_2, \ldots, n_{4^k})$. The correlation $C(A, B)$ between any two species A and B is calculated as the cosine function of the angle between the two vectors:

$$C(A, B) = \frac{\sum_{i=1}^{4^k}(m_i \times n_i)}{\left(\sum_{i=1}^{4^k} m_i^2 \times \sum_{i=1}^{4^k} n_i^2\right)^{\frac{1}{2}}}. \tag{4}$$

The distance $dis(A, B) = \frac{1-C(A,B)}{2}$. Obviously, the distance is normalized to the interval [0,1]. The smaller is the distance, the closer is two species A and B.

3 Results and discussions

3.1 Phylogenetic trees for 48 HEV (Hepatitis E virus)

In the previous study, many efforts have been made to analyze the phylogenetic tree of Hepatitis E virus [24–27]. In our present work, we select the whole genome sequences of 48 HEV (Hepatitis E virus) (Table 1)to analyze the phylogenetic relationship. 48 HEV genomes were distinctly divided into four genotypes. Genotype I included 16 HEV strains. Seven strains, B1 (Bur-82), B2 (Bur-86), I2 [Mad-93], I3, NP1

Table 1 48 Hepatitis E viruses, accession numbers, length and country

No.	Strain name	Accession no.	Length	Genotype	Country
1	B1(Bur-82)	M73218	7,207	I	Burma(Rangoon)
2	B2(Bur-86)	D10330	7,194	I	Burma(Rangoon)
3	I2[Mad-93]	X99441	7,194	I	India(Madras)
4	I3	AF076239	7,194	I	India(Hyderabad)
5	NP1(TK15/92)	AF051830	7,199	I	Nepal(Kathamandu)
6	P2[Abb-2B]	AF185822	7,143	I	Pakistan(Abbotabad)
7	Yam-67	AF459438	7,206	I	India(Yamuna Nagar)
8	C1(CHT-88)	D11092	7,207	I	China(Xinjiang,Hetian)
9	C2(KS2-87)	L25595	7,221	I	China(Xinjiang,Kashi)
10	C3(CHT-87)	L08816	7,176	I	China(Xinjiang,Hetian)
11	C4(Uigh179)	D11093	7,194	I	China(Xinjiang,Uighur)
12	China hebei	M94177	7,200	I	China(Hebei)
13	P1(Sar-55)	M80581	7,138	I	Pakistan(Rangoon)
14	I1(FHF)	X98292	7,202	I	India
15	Morocco	AY230202	7,212	I	Morocco
16	T3	AY204877	7,170	I	Chad
17	M1	M74506	7,180	II	Mexico(Telixtac)
18	HE-JA10	AB089824	7,262	III	Japan(Tokyo)
19	JKN-Sap	AB074918	7,256	III	Japan(Sapporo)
20	JMY-HAW	AB074920	7,240	III	Japan(Sapporo)
21	SW-US1	AF082843	7,207	III	USA
22	US1	AF060668	7,202	III	USA(Minnesota)
23	US2	AF060669	7,277	III	USA(Tennessee)
24	JBOAR1-HYO04	AB189070	7,247	III	Japan(Hyogo)
25	JDEER-HYO03L	AB189071	7,230	III	Japan(Hyogo)
26	JJT-KAN	AB091394	7,218	III	Japan(Kanagawa)
27	JMO-HYO03L	AB189072	7,180	III	Japan(Hyogo)
28	JRA1	AP003430	7,230	III	Japan(Tokyo)
29	JSO-HYO03L	AB189073	7,180	III	Japan(Tokyo)
30	JTH-HYO03L	AB189074	7,180	III	Japan(Tokyo)
31	JYO-HYO03L	AB189075	7,180	III	Japan(Tokyo)
32	SWJ570	AB073912	7,257	III	Japan(Tochigi)
33	KYRGYZ	AF455784	7,239	III	Kyrgyzstan
34	ARKELL	AY115488	7,255	III	Canada(Ontario,Guelph)
35	HE-JA1	AB097812	7,258	IV	Japan(Hokkaido)
36	HE-JK4	AB099347	7,250	IV	Japan(Tochigi)
37	HE-JI4	AB080575	7,186	IV	Japan(Tochigi)
38	JAK-Sai	AB074915	7,236	IV	Japan(Saitama)
39	JKK-SAP	AB074917	7,235	IV	Japan(Sapporo)
40	JSM-SAP95	AB161717	7,202	IV	Japan(Hokkaido)

Table 1 continued

No.	Strain Name	Accession No.	Length	Genotype	Country
41	JSN-SAP-FH	AB091395	7,234	IV	Japan(Hokkaido)
42	JSN-SAP-FH02C	AB200239	7,251	IV	Japan(Hokkaido)
43	JTS-SAP02	AB161718	7,202	IV	Japan(Hokkaido)
44	JYW-SAP02	AB161719	7,202	IV	Japan(Hokkaido)
45	SWJ13-1	AB097811	7,258	IV	China(Uighur)
46	SWCH25	AY594199	7,270	IV	China(Beijing)
47	T1	AJ272108	7,232	IV	USA
48	CCC220	AB108537	7,193	IV	China(Changchun)

(TK15/ 92), P2 [Abb-2B], and Yam-67 were classified into subtype Ia. Subtype Ib contained only I1 (FHF) strain that was isolated from India. Six strains, C1, C2, C3, C4, China hebei and P1 (Sar-55) were clustered together and classified into subtype Ic. Subtype Id and subtype Ie were represented by Morocco strain, and T3 strain from Chad, respectively. In some studies, Morocco and T3 were categorized into a single African cluster, namely genotype V [26,27]. Genotype II contained only a complete genome M1, which was isolated from Mexico. Genotype III included 17 HEV strains. Among these strains, seven strains (HE-JA10, JMY-HAW, JKN-Sap, SWUS1, US1, US2, ARKELL) were classified into subtype IIIa. They were derived from Japan (HE-JA10, JMY-HAW, JKN-Sap), USA (SWUS1, US1, US2), and Canada (ARKELL), respectively. JSO-HYO03L, JMO-HYO03L, JYO-HYO03L, JTH-HYO03L, JDEER-HYO03L, JBOAR1-HYO04, JRA1, JJT-KAN, and SWJ570 were classified into IIIb. They were all derived from Japan. KYRGYZ, which was derived from Kyrgyzstan, was classified into IIIc. Genotype IV included 14 HEV strains. JAK-Sai, JSM-SAP95, JTS-SAP02, JYW-SAP02, JKK-SAP, HE-JI4, SWJ13-1, JSN-SAP-FH, JSN-SAP-FH02C, HE-JK4 and HE-JA1 were classified into genotype IVa. They were all derived from Japan. T1 and SWCH25 were classified into genotype IVb. CCC220 was classified into genotype IVc. T1, SWCH25 and CCC220 were derived from Beijing, Xinjiang and Changchun, respectively.

According to equations in Sects. 2.1 and 2.2, we calculate the distance between arbitrary two sequences and derive a pair-wise distance matrix. This distance matrix contains the similarity information on the 48 HEV. We put the pair-wise distance matrix into the neighbor-joining program in the PHYLIP package (choosing the UPGMA method) [28].

We calculate the similarity matrices for all $k = 2, 3, 4, 5, 6, 7, 8, 9$. In Fig. 1, we outline the phylogentic trees at word lengths $k = 5, 6$. The result of experiments shows that the trees reconstructed from our method converge with k increasing. When $k = 4$ (not shown in this paper), at the overall level, four genotypes of 48 HEV is not separated clearly. However, when $k = 5$, the division of 48 HEV into four genotypes is a clean and prominent feature. And when $k = 6$, different strains of the same genotype come together as they should. When $k = 7, 8, 9$ (not shown in this paper), all genotypes and all different strains are also separated clearly. In other words, the

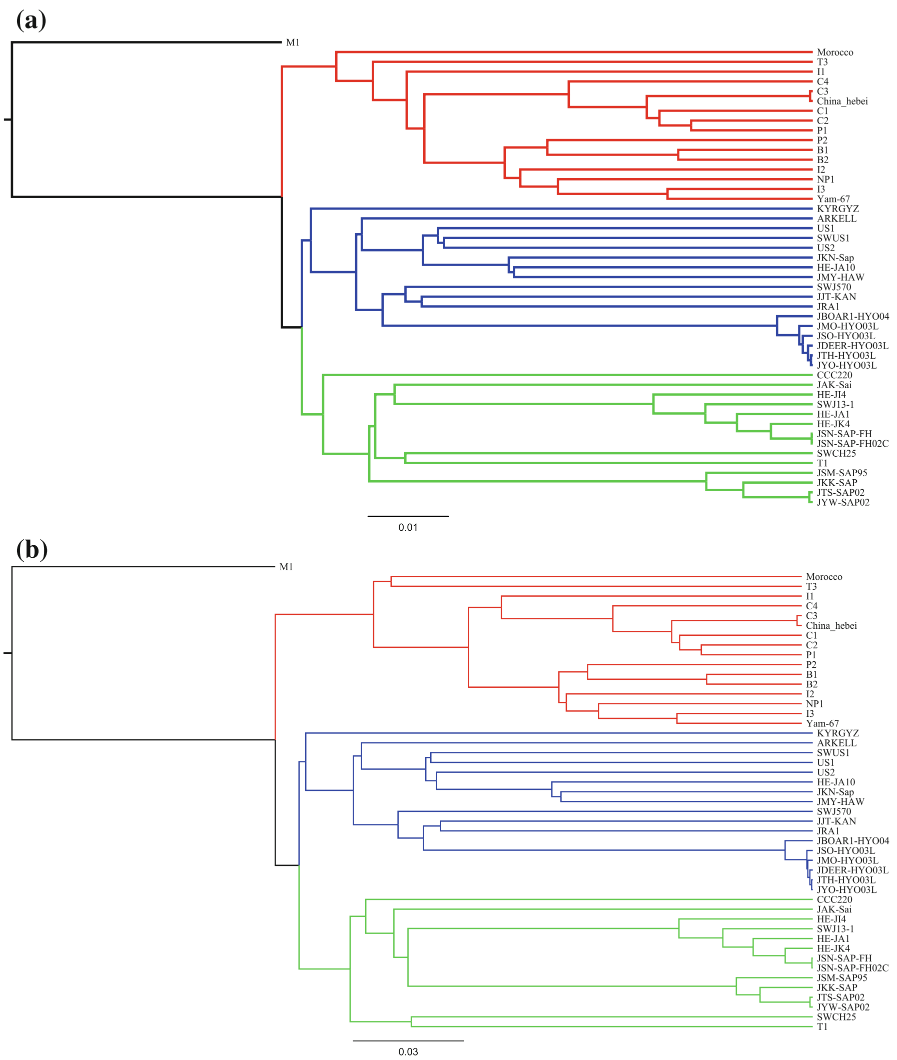

Fig. 1 Phylogenetic trees of 48 HEV for different values for k by our method (for $k = 5$ and 6, the respective trees are shown as (**a**) and (**b**))

topology of the phylogenetic trees became stable from $k = 5$, which can be regarded as a turning point. When $k = 6$ and 7,8,9, the phylogenetic trees are good agreement with that of previous studies [24,25].

3.2 Phylogenetic trees for 20 eutherian mammals

In order to further validate our method, we use the mtDNA sequences of 20 Eutherian mammals as our second data set: human (Homo sapiens, V00662), common chimpanzee (Pan troglodytes, D38116), pigmy chimpanzee (Pan paniscus, D38113), gorilla (Gorilla gorilla, D38114), orangutan (Pongo pygmaeus, D38115), gibbon (Hylobates lar, X99256), baboon (Papio hamadryas, Y18001), horse (Equus caballus, X79547), white rhinoceros (Ceratotherium simum, Y07726), harbor seal (Phoca vitulina, X63726), gray seal (Halichoerus grypus, X72004), cat (Felis catus, U20753), fin whale (Balenoptera physalus, X61145), blue whale (Balenoptera musculus, X72204), cow (Bos taurus, V00654), rat (Rattus norvegicus, X14848), mouse (Mus musculus, V00711), opossum (Didelphis virginiana, Z29573), wallaroo (Macropus robustus, Y10524) and platypus (Ornithorhyncus anatinus, X83427). This dataset consists of seven Primates, eight Ferungulates, two Rodents and three non-placental mammals.

In this experiment, we also calculate the similarity matrices for all $k = 2, 3, 4, 5, 6, 7, 8, 9$. Similar to the first data set, the phylogenetic trees reconstructed from our method converge with k increasing. When $k = 3, 4, 5$, we have no good classifications of Eutherian mammals. When $k = 5$, for seven Primates, eight Ferungulates, two Rodents and three non-placental mammals, different species of the same genus can not come together as they should. However, when $k = 6$, different species of seven Primates, eight Ferungulates and two Rodents come together as they should. But the wallaroo is out of the strain. Nevertheless, when $k = 7$, three main groups of placental mammals, namely Primates, Ferungulates and Rodents, cluster accordingly, and tree non-placental mammals stay outside of all other species. This topology is almost in consistent with some results that given by Arnason [29], Reyes et al. [30], Prasad, Arjun B. et al. [31], Zheng [32], Otu and Sayood [33]. In Fig. 2, we outline the phylogentic trees at word lengths $k = 6, 7$.

In fact, if we reconstruct the the phylogentic trees by NJ method, there are also reliable results. In Fig. 3, we list phylogentic trees of above two data sets in the case $k = 7$.

Further, simulated sequences by means of the software INDELible [34] are applied to evaluate the power and robustness of the proposed method. Sequence data are simulated under the JC model. The number of replicate data sets is 10. The number of sequences in each set is 8 and the length of a sequence is 4,000. The guide tree is shown in Fig. 4. Performance is evaluated by comparing returned trees to the guide tree using symmetric difference distance of Robinson and Foulds (i.e. RF distance) [35] between trees. The distance values are listed in Table 2. From this table, we can see that the values RF distance are smaller ones, which indicates that our method is powerful and robust for evolution analysis of DNA sequences.

3.3 Analysis of adjusting the background probability

In Eq. (1), we define a probability distribution of a k-word and simultaneously consider the subtraction of random background. Finally, we obtain a 4^k-dimensional vector to characterize a DNA sequence. What is the impact of the subtraction of random background? Here, we analyze the differences between these two cases with sample standard deviation.

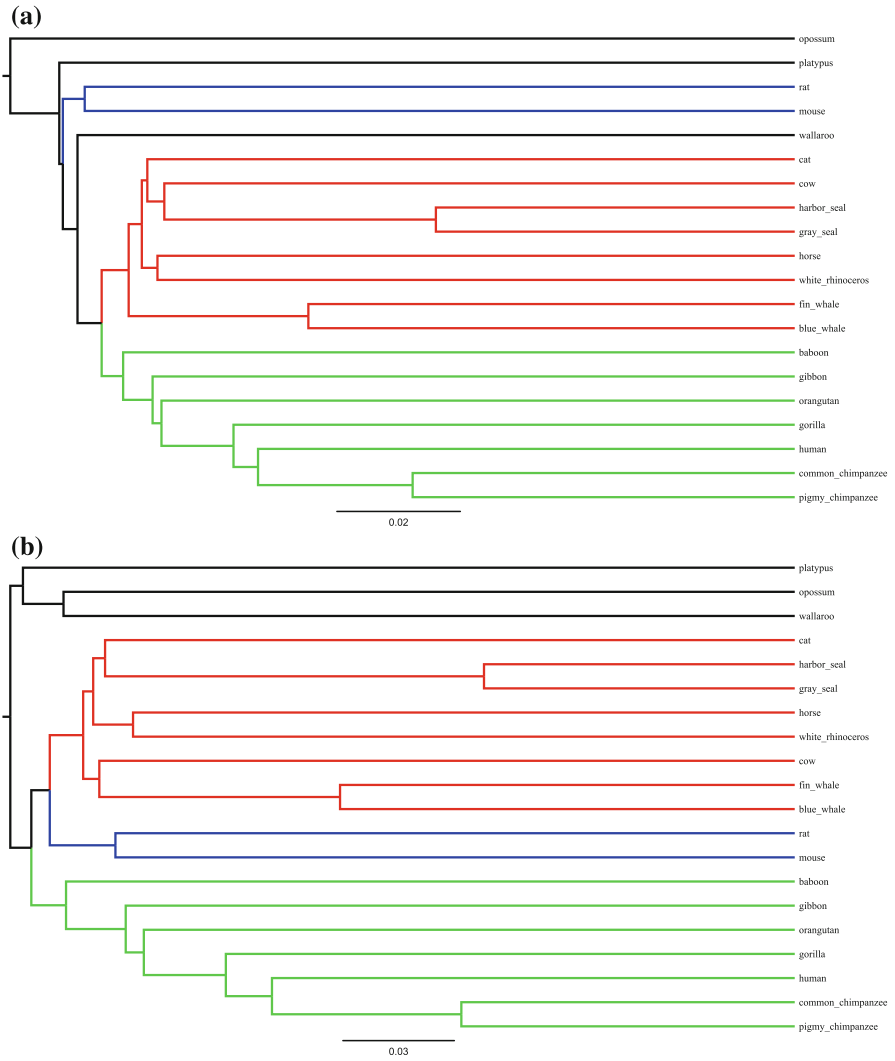

Fig. 2 Phylogenetic trees of 20 Eutherian mammals at $k = 6$ and 7 by our method (the respective trees are shown as (**a**) and (**b**))

Given n samples, every sample has m indicators whose observation data is denoted by x_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$. Putting these data into a matrix, we can get an $n \times m$ matrix denoted by D, i.e.

$$D = \begin{pmatrix} x_{11}, x_{12}, \ldots, x_{1m} \\ x_{21}, x_{22}, \ldots, x_{2m} \\ \cdots \quad \cdots \\ x_{n1}, x_{n2}, \ldots, x_{nm} \end{pmatrix} \tag{5}$$

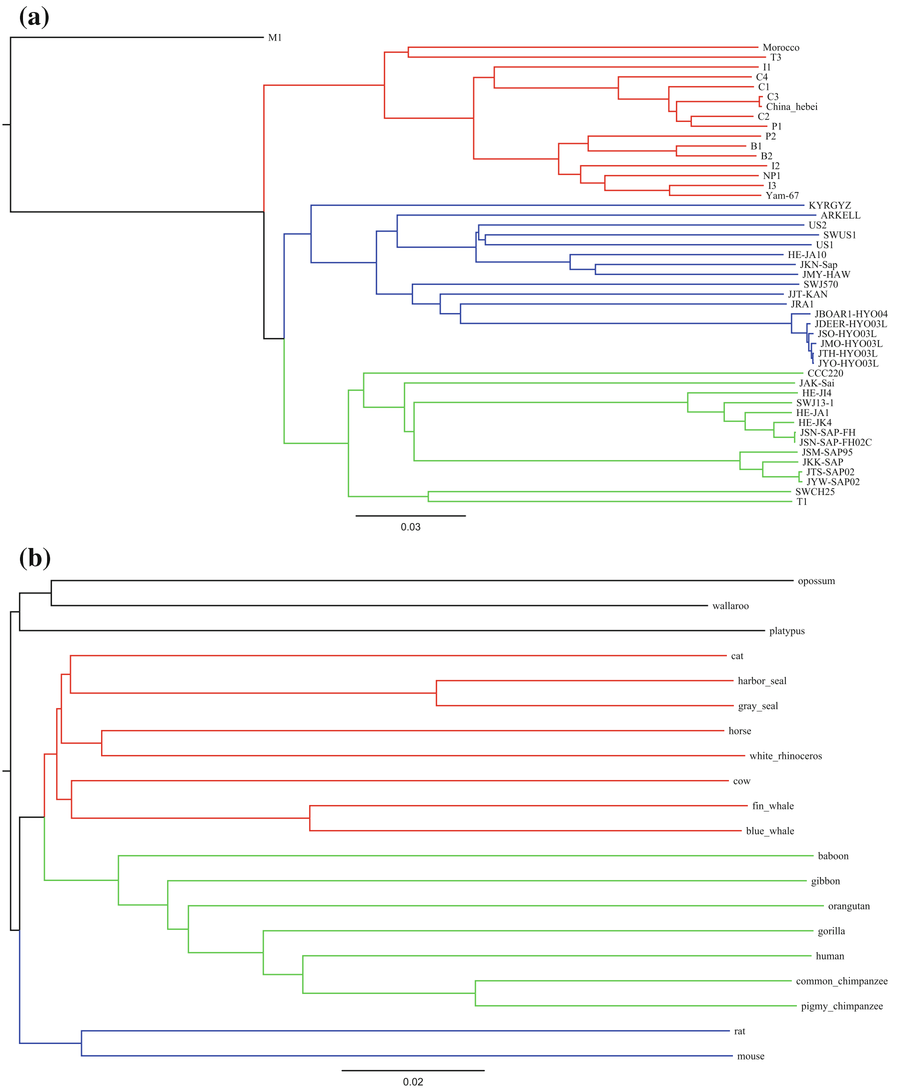

Fig. 3 Phylogenetic trees of 48 HEV and 20 Eutherian mammals by our distance matrix with Neighbor-Joining method

Next, we can calculate every column's standard deviation

$$S_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}, \quad \bar{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \; j = 1, 2, \ldots, m. \tag{6}$$

Then, we average the set of $\{S_j\}$, $j = 1, 2, \ldots, m$, and get its mean denoted by S. The higher is the value of S, the greater is the degree of dispersion of the data, which

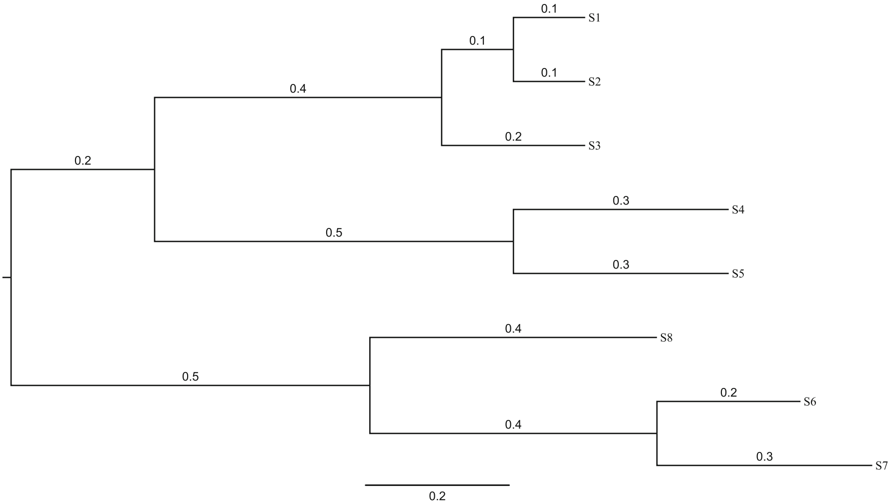

Fig. 4 Phylogenetic tree of 8 simulated sequences which is the guide tree. The evolutionary distance are shown on the branches of the tree

indicates that the data involves the richer data information. Therefore, we can take the S as the indicator to measure the divergence of the two different methods.

If we collect the maximum of the probability of each k-word's distribution without considering the background probability, we will obtain a 4^k-dimensional vector for each DNA sequence (Case I). After subtracting the background probability, we also derive a 4^k-dimensional vector for each DNA sequence according to the Eq. (3) (Case II).

Next, we take the first data set i.e. 48 HEV (Hepatitis E virus) for an example to elucidate the difference of the two above cases. Interpreting n samples as 48 HEV and m indicators as $4^k k$-words, for a fixed k, each sequence of 48 HEV corresponds a 4^k-dimensional characteristic vector. Thereby, we have a matrix D_k and a standard deviation S_k according to (5) and (6). In Table 3, we list the values of S_k of Case I and Case II for 48 HEV and the corresponding curves are displayed in Fig. 5a. Table 2 or Fig. 5a denotes that for each k, the value of S_k of Case II is almost more than or equal that of Case I, which indicates that the degree of dispersion of the data in Case II is greater than that in Case I. Therefore, considering the background probability contains more richer sequence information. In addition, we find that the value of S_k is increasing with k from 2 to 7 but decreasing at $k = 8$. When the length of the k-word is long, the number of occurrence in the DNA sequence is small, which results in the matrix D_k is a sparse matrix. Therefore, the corresponding standard deviation S_k becomes smaller. So we can surmise that k-words whose length are 5,6,7 and 8 involve more plentiful information. In our experiments, we get the perfect phylogentic tree at $k = 6$ for 48 HEV. In the experiment of the second data set, we also do above work and get the similar conclusion (Fig. 5b). We get the perfect phylogentic tree at $k = 7$ for 20 Eutherian mammals. However, if we don't consider the random background,

Table 2 The values of RF distance between the guide tree and the returned tree

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Guide tree	2	2	6	2	4	2	2	6	4	6

T*i* denotes the *i*th returned tree

Table 3 The values of S_k of Case I and Case II for 48 HEV

k	2	3	4	5	6	7	8	9
Case I	0.0026	0.0110	0.0361	0.1337	0.3287	0.3414	0.1799	0.0645
Case II	0.0074	0.0458	0.1149	0.1917	0.3385	0.3430	0.1800	0.0644

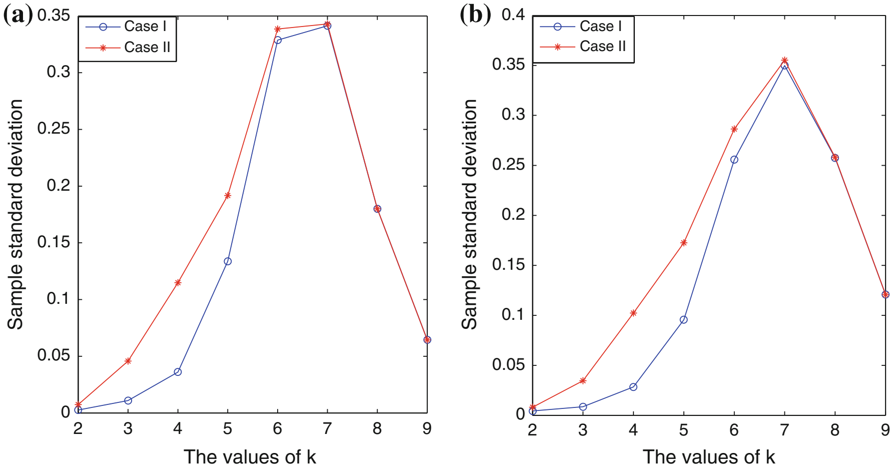

Fig. 5 Sample standard deviation S_k of Case I (without considering the random background) and Case II (our method) for different values of *k* from (**a**) 48 HEV, (**b**) 20 Eutherian mammals

Table 4 The turning point of *k* of the perfect phylogentic tree for different methods

	Case I	Case II (our method)	CV method
48 HEV	$k = 7$	$k = 5$	$k = 7$
20 Euth mammals	none	$k = 6$	$k = 7$

the perfect phylogentic tree is obtained at $k = 7$ for 48 HEV, but none for 20 Eutherian mammals.

In this section, for the reconstruction of phylogentic tree of the two data sets, we not only compare the difference of Case I and Case II but also compare the two cases with CV method. From Table 4, we can see that for the two data sets, the values of *k* obtained by our method is always less than that of other two methods (Case I and CV method). These results indicate that our method capture more sequence information so that we can construct the perfect phylogentic trees at the small value of *k* and avoid the long running time. Finally we conclude that it is necessary to consider both the position and number of a k-word and the background probability. In Table 4, we list the turning point of k for different methods and data sets.

4 Conclusions

Due to the computational time and the inherent model assumptions, multiple sequence alignment of genomic sequences is a bottleneck. So, there is a great need to develop new sequence comparisons free of these problems, especially for whole genomic sequences.

In this paper, we propose a new alignment-free method to analyze the evolutionary relationship of the genomic sequences. we define a probability distribution of a k-word and simultaneously consider the subtraction of random background. Finally, we derive a 4^k-dimensional vector to characterize a DNA sequence. In order to capture more biological information, we take both the position and number of a k-word and the random background into account. Through reconstructing the phylogenetic tree of 48 HEV and the mtDNA sequences of 20 Eutherian mammals with our method, we find that the results are good consisted with classical ones, which indicates that our method is an useful and effective tool to phylogenetic analysis. To further evaluate our method, we compare it with Case I (without considering the random background) and CV method (Fig. 5 and Table 4). The comparison demonstrates that our method owns more sequence information and is more efficient to perform the phylogenetic analysis.

However, our method has its own disadvantages and many improvements remain to be discovered and developed. Because we consider not only the numbers of the occurrence of the k-word but also its positions. The time-consuming of our algorithm is not superior to other methods but does not run very slowly. The time complexity is $4^k * O(n)$, where k is the length of a k-word and n is the number of a k-word occurring in a DNA sequence. Our method have relatively high computational cost which is more than an hour when $k \geq 9$. In our future study, we will further improve our model and algorithm to overcome this problem.

Acknowledgments This work is supported by the National Natural Science Foundation of China(Grant No.10871219), Natural Science Foundation of Shandong Province (Grant No.ZR2010GL013), Humanities and Social Sciences Project of the Ministry Education of China (Grant No.09YJC910004; Grant No.10YJC630092), Research Program of Higher Education of Shandong Province (Grant No.J10WF84) and Natural Science Foundation of Shandong Jiaotong University (Grant No.Z201031).

References

1. M. S. Waterman, *Introduction to computational biology: maps, sequeces, and genomes* (Chapman & Hall, New York, 1995)
2. R. Durbin, S.R. Eddy, A. Krogh, G. Mitchison, *Biological sequence analysis* (Cambridge University Press, Cambridge, 1998)
3. M. Randic, M. Vracko, J. Chem. Inf. Comput. Sci. **40**, 599 (2000)
4. M. Randic, M. Vracko, N. Lers, D. Plavsic, Chem. Phys. Lett. **368**, 1 (2003)
5. M. Randic, A.T. Balaban, J. Chem. Inf. Comput. Sci. **43**, 532 (2003)
6. B. Liao, T.M. Wang, Chem. Phys. Lett. **388**(1-3), 195 (2004)

7. G.H. Huang, B. Liao, Y.F. Li, Y.G. Yu, Biophys. Chem. **143**, 55 (2009)
8. B. Liao, Chem. Phys. Lett. **401**, 196 (2005)
9. J. Berger, S. Mitra, M. Carli, A. Neri, J. Franklin Inst. **341**, 37 (2004)
10. Y.H. Yao, T.M. Wang, Chem. Phys. Lett. **398**, 318 (2004)
11. Y.H. Yao, X.Y. Nan, T.M. Wang, Chem. Phys. Lett. **411**, 248 (2005)
12. S. Vinga, J. Almeida, Bioinformatics **19**, 513 (2003)
13. G. Reinert, S. Schbath, M.S. Waterman, J. Comput. Biol. **7**, 1 (2000)
14. Q. Dai, T.M. Wang, Bioinformatics **24**, 2296 (2008)
15. Y.J. Huang, L.P. Yang, T.M. Wang, J. Theor. Biol. **269**(1), 217 (2011)
16. B.E. Blaisdell, *Proc. Natl. Acad. Sci. USA.* **83**,5155 (1986)
17. T.J. Wu, J.P. Burke, D.B. Davison, Biometrics **53**, 1431 (1997)
18. T.J. Wu, Y.C. Hsieh, L.A. Li, Biometrics **57**, 441 (2001)
19. G.W. Stuart, K. Moffect, S. Baker, Bioinformatics **18**, 100 (2002)
20. B.L. Hao, J. Qi, J. Bioinf. Comput. Biol. **2**, 1 (2004)
21. L. Gao, J. Qi, B.L. Hao, AAPPS Bull. **6**, 3 (2006)
22. J. Qi, B. Wang, B.L. Hao, J. Mol. Biol. **58**, 1 (2004)
23. H. Wang, Z. Xu, L. Gao, B.L. Hao, BMC Evol. Biol. **9**, 195 (2009)
24. L. Lu, C. Li, C.H. Hagedorn, Rev. Med. Virol. **16**, 5 (2006)
25. Z.H. Liu, J.H. Meng, X. Sun, Biochem. Biophys. Res. Commun. **368**, 223 (2008)
26. R. Chatterjee, S. Tsarev, J. Pillot, P. Coursaget, S.U. Emerson, R.H. Purcell, J. Med. Virol. **53**, 139 (1997)
27. H. van Cuyck-Gandre, H.Y. Zhang, S.A. Tsarev, N.J. Clements, S.J. Cohen, J.D. Caudill, Y. Buisson, P. Coursaget, R.L. Warren, C.F. Longer, J. Med. Virol. **53**, 340 (1997)
28. J. Felsenstein, *PHYLIP (Phylogenetic Inference Package) ver. 3.57.* (Department of Genetics, University of Washington, Seattle, WA, 1995)
29. U. Arnason, J.A. Adegoke, K. Bodin, E.W. Born, Y.B. Esa, A. Gullberg, M. Nilsson, R.V. Short, X.f. Xu, A. Janke, Proc. Natl. Acad. Sci. USA. **99**(12), 8151 (2002)
30. A. Reyes, C. Gissi, F. Catzeflis, E. Nevo, G. Pesole, C. Saccone, Mol. Biol. Evol. **21**(2), 397 (2004)
31. A.B. Prasad, M.W. Allard, E.D. Green, Mol. Biol. Evol. **25**(9), 1795 (2008)
32. X.Q. Zheng, Y.F. Qin, J. Wang, Math. Biosci. **217**, 159 (2009)
33. H.H. Otu, K. Sayood, Bioinformatics **19**, 2122 (2003)
34. W. Fletcher, Z.H. Yang, Mol. Biol. Evol. **26**(8), 1879 (2009)
35. D. Robinson, L. Foulds, Math. Biosci. **53**, 131 (1981)