Mathematical Biosciences 217 (2009) 159–166

Contents lists available at ScienceDirect

Mathematical Biosciences

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

A Poisson model of sequence comparison and its application to coronavirus phylogeny

Xiaoqi Zheng [a,b], Yufang Qin [a], Jun Wang [c,*]

[a] Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, PR China
[b] College of Advanced Science and Technology, Dalian University of Technology, Dalian 116024, PR China
[c] Department of Mathematics, Shanghai Normal University, Shanghai 200034, PR China

ARTICLE INFO

Article history:
Received 23 January 2008
Received in revised form 30 September 2008
Accepted 14 November 2008
Available online 6 December 2008

Keywords:
Word composition
Poisson model
Similarity
Coronavirus phylogeny

ABSTRACT

In this paper, we propose two metrics to compare DNA and protein sequences based on a Poisson model of word occurrences. Instead of comparing the frequencies of all fixed-length words in two sequences, we consider (1) the probability of 'generating' one sequence under the Poisson model estimated from the other; (2) their different expression levels of words. Phylogenetic trees of 25 viruses including SARS-CoVs are constructed to illustrate our approach.

1. Introduction

One of the fundamental tasks in bioinformatics is sequence comparison, which is used heavily in database searching, sequence classification, phylogenetic tree reconstruction and detection of regulatory sequences. In most cases, alignments are performed between the target sequences by dynamic programming techniques and the resulting alignment scores are used to calculate a measure of similarity. Meanwhile, especially in recent years, an increasing number of alignment-free methods have emerged [1–4]. In contrast to traditional alignments, these alignment-free methods mostly (i) make few assumptions of the evolutionary model and (ii) present light computational load. With the first merit, alignment-free methods do not suffer greatly from some evolutionary events, e.g., large rearrangements and transposon activity. While the second merit enables broad contributions of alignment-free comparisons in pre-filtering relevant sequences, and then using alignment algorithms to refine the searches. This type of heuristic approach is already used in programs like BLAST [5] and FASTA [6]. Additionally, after the completion of many genome projects, alignment-free comparisons begin to find their use in whole genome phylogeny, which meets great computational and theoretical challenges using alignment-based methods.

Sequence comparison based on word statistics may be the most well-developed alignment-free method. Observing that relative abundances of all dinucleotides are remarkably constant across the genome, Karlin et al. [7–9] proposed the 'genome signature' to describe a genome. The 'signature' consists of the array of dinucleotide relative abundances $\rho_{xy} = f_{xy}/f_x f_y$ extended over all dinucleotides, where f_x is the frequency of nucleotide x and f_{xy} is the frequency of dinucleotide xy. In the same manner, genome signature based on abundances of k-nucleotides can also be defined. Reinert et al. [10] studied the statistical and probabilistic properties of words in sequences, with emphasis on the deductions of exact distributions and evaluation of its asymptotic approximations. Word-based comparisons were recently reviewed by Vinga and Almeida [2]. According to their work, biological sequences are first represented as frequency vectors in Euclidean space, and then pairwise distances between these sequences can be defined as the standard Euclidean distance, Mahalanobis distance, linear correlation coefficient or Kullback–Leibler discrepancy between their corresponding vectors. As another powerful tool for sequence analysis, some graphical representations of DNA or protein sequences are also based on statistics of short words [11,12].

In this paper, we propose two distance measures for biological sequences on the basis of word statistics. Instead of comparing the frequencies or relative compositions of each word type in two sequences, we explore two measures in the probabilistic framework. Some basic concepts and our computational methods are introduced in the following section. To illustrate our method,

* Corresponding author. Tel.: +86 411 8470 6101; fax: +86 411 8470 6100.
E-mail address: junwang@dlut.edu.cn (J. Wang).

0025-5564/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.mbs.2008.11.006

in Section 3, similarity trees of 25 virus genomes are built by some classical distances and our methods.

2. Methods

A sequence S, of length l, is defined as a linear succession of symbols from a finite alphabet $\mathscr{A}$, of length n. A k-word (or k-mer, k-tuple, etc.) $\omega = \alpha_1\alpha_2\cdots\alpha_k$ is a subsequence of k adjacent letters, $\alpha_i \in \mathscr{A}$, $i = 1, 2, \ldots, k$. Obviously, there are a total of n^k possible k-words for the alphabet $\mathscr{A}$. The occurrence of ω (denoted by N_ω) is the number of times it is seen through sliding a window of width k once across the sequence, and frequency of this word f_ω is obtained by simply dividing the total number of words (i.e., $f_\omega = N_\omega/(l-k+1)$). Given a symbol sequence, we can represent it as a point in the high dimensional Euclidean space by a mapping from S to the vector of its word counts, or frequencies:

$$N(S) = (N_{\omega_1}, N_{\omega_2}, \ldots, N_{\omega_{n^k}}) \quad \text{or} \quad f(S) = (f_{\omega_1}, f_{\omega_2}, \ldots, f_{\omega_{n^k}}).$$

For DNA sequences, $\mathscr{A} = \{A, G, C, T\}$. If 2-words are considered and words in above vectors are arranged as (AA, AG, AC, AT, GA, GG, ..., TT), the corresponding vectors for S = AAAGGA are

$$N(S) = (2, 1, 0, 0, 1, 1, 0, \ldots, 0) \quad \text{and}$$
$$f(S) = (0.4, 0.2, 0, 0, 0.2, 0.2, 0, \ldots, 0).$$

To evaluate the distance between two sequences, it is intuitive to compute the norm of the difference between their corresponding frequency (or occurrence) vectors,

$$d(S_1, S_2) = \sqrt[p]{\sum_{i=1}^{n^k} |f_{1,\omega_i} - f_{2,\omega_i}|^p},$$

where f_{1,ω_i} and f_{2,ω_i} are frequencies of the word ω_i in sequences S_1 and S_2, respectively. The norm gives mathematically well defined distance functions for all positive values of p. Here $p = 1$ gives the Manhattan distance, which was used in [7,13]; $p = 2$ gives the Euclidean distance [14]; $p = \infty$ gives the max-norm (where only the largest absolute value contributes). However, these simple distances are not satisfying for an accuracy phylogeny, because (i) they treat all word types equally, despite that they have different background frequencies, and (ii) contribution of a word may not merely be a polynomial function of the frequency difference. In order to overcome the above problems, the Mahalanobis and standard Euclidean distance, which take into account the data covariance structure, were proposed for sequence comparison relatively recently [15]. In this paper, we will propose two distance measurements free of such problems by using a probabilistic framework.

The most immediate model for word occurrences is the binomial distribution, i.e., each word ω has the same probability p to appear at any word location. When p is very small, sequence length l is sufficiently large, and the value of lp is moderate, the occurrences of ω in this sequence approximately follow the Poisson distribution with the parameter lp. In what follows we will explore two distance metrics on the basis of the Poisson distribution of word occurrences.

2.1. The relative Poisson distance

For simplicity, we assume that S_1 and S_2 have the same length l (or else we can normalize one of them). Occurrences of word ω_i in these two sequences are denoted by N_{1,ω_i} and N_{2,ω_i}, respectively. In the first step, we use S_1 to estimate the Poisson parameter. Known that the parameter λ of Poisson model is equal to the expectation of the variable (word occurrence, in our model), we intuitively set $\lambda = N_{1,\omega_i}$. Then define

$$\mathrm{RP}_{\omega_i}(S_1, S_2) = \mathrm{Poi}(N_{2,\omega_i}; N_{1,\omega_i}) = \frac{(N_{1,\omega_i})^{N_{2,\omega_i}} \cdot e^{-N_{1,\omega_i}}}{N_{2,\omega_i}!},$$

where $\mathrm{Poi}(k;\lambda)$ is the Poisson probability with parameter λ,

$$\mathrm{Poi}(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{1}$$

Actually, $\mathrm{RP}_{\omega_i}(S_1, S_2)$ measures a kind of 'similarity' between S_1 and S_2 in terms of the occurrences of ω_i (note that it is not a strict similarity measure as it is not symmetrical). Explicitly, low values of RP_{ω_i} correspond to the relatively large discrepancies in occurrences of the word ω_i, and the maximum value is gotten when $N_{1,\omega_i} = N_{2,\omega_i}$ or $N_{1,\omega_i} = N_{2,\omega_i} + 1$. Taking all words into consideration, the final distance between S_1 and S_2 is defined

$$d_{\mathrm{RP}}(S_1, S_2) = \sum_{i=1}^{n^k} (\mathrm{RP}_{\omega_i}(S_1, S_1) + \mathrm{RP}_{\omega_i}(S_2, S_2) - \mathrm{RP}_{\omega_i}(S_1, S_2) - \mathrm{RP}_{\omega_i}(S_2, S_1)). \tag{2}$$

Here the two terms $\mathrm{RP}_{\omega_i}(S_1, S_1)$ and $\mathrm{RP}_{\omega_i}(S_2, S_2)$ are introduced to guarantee the positivity of $d_{\mathrm{RP}}(S_1, S_2)$ (note that $\mathrm{RP}_{\omega_i}(S_1, S_1) \geqslant \mathrm{RP}_{\omega_i}(S_1, S_2)$ for any word ω_i).

Since $\mathrm{RP}_{\omega_i}(S_1, S_2)$ measures the probability to observe N_{2,ω_i} times of ω_i in sequence S_2 in the condition that the average occurrence is N_{1,ω_i}, we refer to d_{RP} as the *Relative Poisson distance* between S_1 and S_2.

2.2. The distance based on expression level of words

In the above subsection, we consider only one Poisson model – parameter of this model is estimated by one sequence, and pairwise similarity is evaluated by the probability of generating the other sequence under this model. In this part, the occurrences of word ω_i in sequences S_1 and S_2 follow two different Poisson distributions (with parameters $\lambda_{1,i}$ and $\lambda_{2,i}$, respectively). Define

$$\mathrm{Exp}_{1,\omega_i} = \sum_{k=0}^{N_{1,\omega_i}} \mathrm{Poi}(k; \lambda_{1,i}), \tag{3}$$

where N_{1,ω_i} is the occurrence of ω_i in S_1. $\mathrm{Exp}_{1,\omega_i}$ is the probability of observing $\leqslant N_{1,\omega_i}$ occurrences of ω_i in sequence S_1. Note that a word is called highly expressed if its observed frequency is more than its expected frequency, and called low expressed otherwise. In this sense, the probability $\mathrm{Exp}_{1,\omega_i}$ measures a level of expression – low value of $\mathrm{Exp}_{1,\omega_i}$ corresponds to low expression of word ω_1, and large value of $\mathrm{Exp}_{1,\omega_i}$ corresponds to high expression of the word ω_1 in sequence S_1. We define the final distance between S_1 and S_2 as

$$d_{\mathrm{Exp}}(S_1, S_2) = \sum_{i=1}^{n^k} |\mathrm{Exp}_{1,\omega_i} - \mathrm{Exp}_{2,\omega_i}|. \tag{4}$$

Now, to compute $d_{\mathrm{Exp}}(S_1, S_2)$ we need to determine $\lambda_{1,i}$ and $\lambda_{2,i}$ for each word in each sequence. Note that the Poisson parameter for a word is actually its expected occurrence, which can be obtained immediately by multiplying the expected frequency (or background frequency) by the total number of words. We now only need to determine the background frequency of each word. To achieve this aim, two approaches are tried. The first approach corresponds to independence of nucleotides in the sequence, i.e., background frequency of the word ω is estimated by the product of the corresponding nucleotide frequencies in this sequence,

$$\bar{f}_\omega = \bar{f}_{\alpha_1\alpha_2\cdots\alpha_k} = f_{\alpha_1} f_{\alpha_2} \cdots f_{\alpha_k}, \tag{5}$$

where f_{α_i} $(i = 1, 2, \ldots, k)$ is the frequency of the letter α_i in this sequence. An alternative method for estimating the background frequency of a word was proposed by Qi et al. [16], who applied a

Markov model of DNA sequences of order $k-2$. The expected frequency of a word is predicted from the probabilities of appropriate shorter subwords

$$\bar{f}_\omega = \bar{f}_{\alpha_1\alpha_2\cdots\alpha_k} = \frac{f_{\alpha_1\alpha_2\cdots\alpha_{k-1}} \cdot f_{\alpha_2\alpha_3\cdots\alpha_k}}{f_{\alpha_2\alpha_3\cdots\alpha_{k-1}}}, \tag{6}$$

where $f_{\alpha_1\alpha_2\cdots\alpha_{k-1}}$ is the frequency of the $(k-1)$-word $\alpha_1\alpha_2\cdots\alpha_{k-1}$ in the corresponding sequence. Then for each background probability estimated by Eqs. (5) and (6), the corresponding Poisson parameter is

$$\lambda_\omega = \bar{f}_\omega \cdot (l-k+1).$$

The distance d_{Exp} has the following properties: (i) $d_{\text{Exp}}(S_1,S_2) \geqslant 0$ and $d_{\text{Exp}}(S_1,S_1)=0$, for any sequences S_1 and S_2; (ii) background information (or frequencies of shorter words) is incorporated into the measurement; and (iii) words with identical frequency and occurrence in two sequences may contribute to d_{Exp}, i.e., they may have different background frequencies and expression levels.

When λ is large ($\lambda > 50$), however, it is difficult to obtain the accurate Poisson probability by Eq. (1) using personal computers. Explicitly, as $e^{-\lambda}$ is very small and λ^k is very large in the numerator, mistakes may be made if they are multiplied directly. In order to overcome this difficulty, another two executive approximations of Poisson probability in the case of large λ are tried: (i) Stirling formula. According to the Stirling formula, $k! \sim (k/e)^k\sqrt{2\pi k}$, so $P(k;\lambda) \doteq \frac{\lambda^k e^{-\lambda}}{(k/e)^k\sqrt{2\pi k}} = (\lambda/k)^k e^{k-\lambda}\sqrt{2\pi k}$. (ii) Normal approximation of Poisson distribution. When λ is sufficiently large, the Poisson distribution with parameter λ can be approximated by the Normal distribution $N(\lambda,\lambda)$.

3. Application

3.1. Phylogenetic trees of 25 viruses including SARS-CoVs

Coronaviruses are the causative agents of a number of mammalian diseases which often have significant economic and health-related consequences [17,18]. On the basis of antigenic cross-reactivity, coronaviruses were originally classified into three groups. Group I and group II contain mammalian viruses (while group II coronaviruses contain a hemagglutinin esterase gene homologous to that of Influenza C virus [19]), and group III contains only avian viruses. After the outbreak of severe acute respiratory syndrome coronavirus (SARS-CoV) in 2003, many efforts have been made to identify the phylogenetic positions of SARS-CoVs in the coronavirus phylogeny. However, this is still a controversial topic – alignment-based methods showed that SARS-CoVs are not closely related to any previously isolated groups and form a new group [20,21]; maximum likelihood tree built from a fragment of the spike protein preferred SARS-CoVs clustering with group II coronaviruses (murine hepatitis virus and rat coronavirus) [22]; while an information-based method, which made use of the whole genome sequences, indicated that the SARS-CoVs should not be classified as a new group but close to the group I coronaviruses [23].

In this paper, we select 25 complete virus genomes: 12 coronaviruses from the three isolated typical groups, 12 SARS-CoV strains, and a torovirus, which serves as the outgroup for coronaviruses [24] (data are shown in Table 1). In order to validate our method, distance matrices for the same data set are also constructed using some classical dissimilarity measurements, e.g., the standard Euclidean distance [15,25], linear correlation coefficient [26], Kullback–Leibler (KL) discrepancy [3] and the Composition Vector approach [16,27]. Note that the Kullback–Leibler discrepancy between two frequency vectors is not symmetrical and will give degenerate results when some word types are absent, we use a revised version – the Weighted Sequence Entropy (WSE) [28]. This modification works equivalently with the KL discrepancy in the case of short words, and can effectively avoid the degeneracy for long words. The string Composition Vector (CV) approach proposed by Hao's group is a fast and efficient approach to whole genome comparison and phylogenetic analysis. For each k-string ω, define

$$\text{CV}_\omega = \begin{cases} \frac{f_\omega - \bar{f}_\omega}{\bar{f}_\omega}, & \bar{f}_\omega \neq 0, \\ 0, & \bar{f}_\omega = 0, \end{cases} \tag{7}$$

where f_ω is the frequency of word ω in a genomic sequence, and $\bar{f}_\omega$ is its expect frequency under a certain background model (Markov model of $k-2$ order). Then collect CV_ω for all possible ω as components to form a composition vector. The final distance between two species is evaluated based on the cosine function between their corresponding composition vectors.

After calculating the pairwise distance matrices, phylogenetic trees for the 25 viruses are built by the UPGMA and NJ programs in the PHYLIP package. Then, rooted phylogenetic trees are drawn by the TREEVIEW program [29]. The UPGMA tree built by the standard Euclidean distance is shown as Fig. 1(1). This tree supports torovirus as the outgroup of all coronaviruses, but fails to cluster three group I coronaviruses – HCoV-229E and PEDV are grouped together, but TGEV is much closer to the SARS clade. Fig. 1(2) is the NJ tree constructed by the Euclidean distance. Similar to the UPGMA tree, this tree also prefers SARS-CoVs clustering with TGEV. But an obvious defect is that it does not successfully cluster the eight group II coronaviruses. In Fig. 2, we list the trees built by linear correlation coefficient between pairwise frequency vectors. Fig. 2(1) is the UPGMA tree. This tree perfectly clusters species within each typical group, and confirmed SARS-CoVs paraphyly. But it fails to identify the outgroup status of torovirus relative to coronaviruses. While the NJ tree (Fig. 2(2)), in which torovirus is selected as outgroup species, confirms the adjacent relationship of SARS-CoVs with group I viruses. In the tree built from our distance measure d_{Exp} (Fig. 3), all above defects are eliminated, i.e., species of each typical groups cluster, and torovirus stays outside of all coronaviruses including SARS-CoVs. Our tree shows that SARS-CoVs are not closely related to any previously isolated coronaviruses and form a new group, but do not support the outgroup status of SARS-CoVs relative to other coronaviruses, as proposed by Zheng et al. [30]. This result is mainly in accordance with the WSE tree at word order $k=6$ (Fig. 4) and the NJ tree constructed by the Composition Vector method (Fig. 5). Moreover, it is also supported by the experimental evidence, which showed that group I coronaviruses specific antibodies are able to recognize antigens in SARS-CoV infected cultured cells [31].

3.2. Whole mitochondrial genome phylogeny of 20 Eutherian mammals

In order to further validate our algorithm, we use the complete mtDNA sequences of 20 Eutherian mammals selected by Otu and Sayood as our second dataset [32]. This dataset consists of seven Primates, eight Ferungulates, two Rodents and three non-placental mammals. Their corresponding GenBank Accession Codes are as follows:

- *Primates*: 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) and baboon (*Papio hamadryas*, Y18001).
- *Ferungulates*: 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) and cow (*Bos taurus*, V00654).

Table 1
Coronaviruses and a torovirus used to constructed phylogenetic tree.

No.	Accession No.	Abbreviation	Genome	Group	Length (nt)
1	NC_002654	HCoV-229E	Human coronavirus 229E	I	27317
2	NC_002306	TGEV	Transmissible gastroenteritis virus	I	28586
3	NC_003436	PEDV	Porcine epidemic diarrhea virus	I	28033
4	U00735	BCoVM	Bovine coronavirus strain Mebuus	II	31032
5	AF391542	BCoVL	Bovine coronavirus isolate BCoV-LUN	II	31028
6	AF220295	BCoVQ	Bovin coronavirus strain Quebec	II	31100
7	NC_003045	BCoV	Bovine coronavirus	II	31028
8	AF208067	MHVM	Murine hepatitis virus strain ML-10	II	31233
9	AF201929	MHV2	Murine hepatitis virus stain 2	II	31276
10	AF208066	MHVP	Murine hepatitis virus stain Penn 97-1	II	31112
11	NC_001846	MHV	Murine hepatitis virus	II	31357
12	NC_001451	IBV	Avian infectious bronchitis virus	III	27608
13	AY278488	BJ01	SARS coronavirus 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-Su10	–	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
25	X52374	EToV	Equine torovirus	–	7920

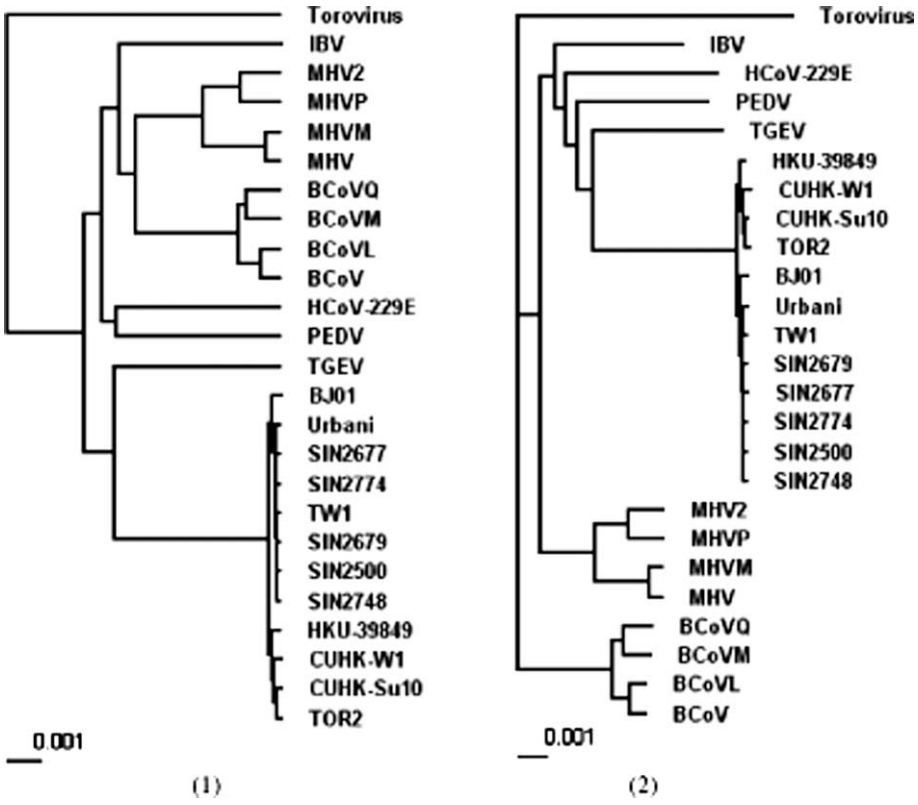

Fig. 1. (1) UPGMA and (2) Neighbor-Joining trees of 25 viruses ($k = 6$). Pairwise distances are evaluated by the standard Euclidean distance.

- *Rodents*: Rat (*Rattus norvegicus*, X14848) and mouse (*Mus musculus*, V00711).
- *Non-placental mammals*: Opossum (*Didelphis virginiana*, Z29573), wallaroo (*Macropus robustus*, Y10524) and platypus (*Ornithorhyncus anatinus*, X83427).

We applied the proposed distance measurements to the complete mitochondrial genomes listed above. In Fig. 6, we list the UPGMA tree constructed by the distance d_{Exp} with background frequencies estimated by Eq. (6). As is seen from this figure, three main groups of placental mammals, namely Primates, Ferungulates and Rodents, cluster accordingly, and three non-placental mammals stay outside of all other species. This topology is in perfect agreement with that given by Otu and Sayood except for the position of rodents (mouse and rat). However, the relationship among the three main groups of placental mammals is still a controversial

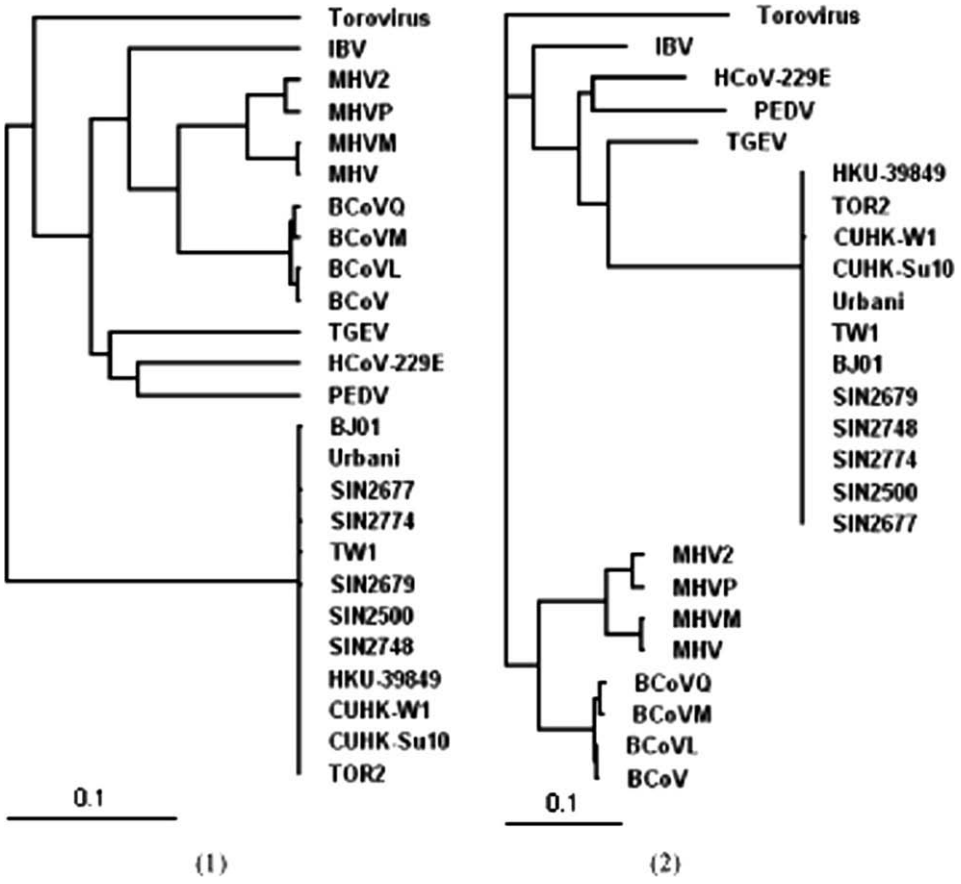

Fig. 2. (1) UPGMA and (2) Neighbor-Joining trees of 25 viruses ($k = 6$). Pairwise distances are evaluated by the linear correlation coefficient.

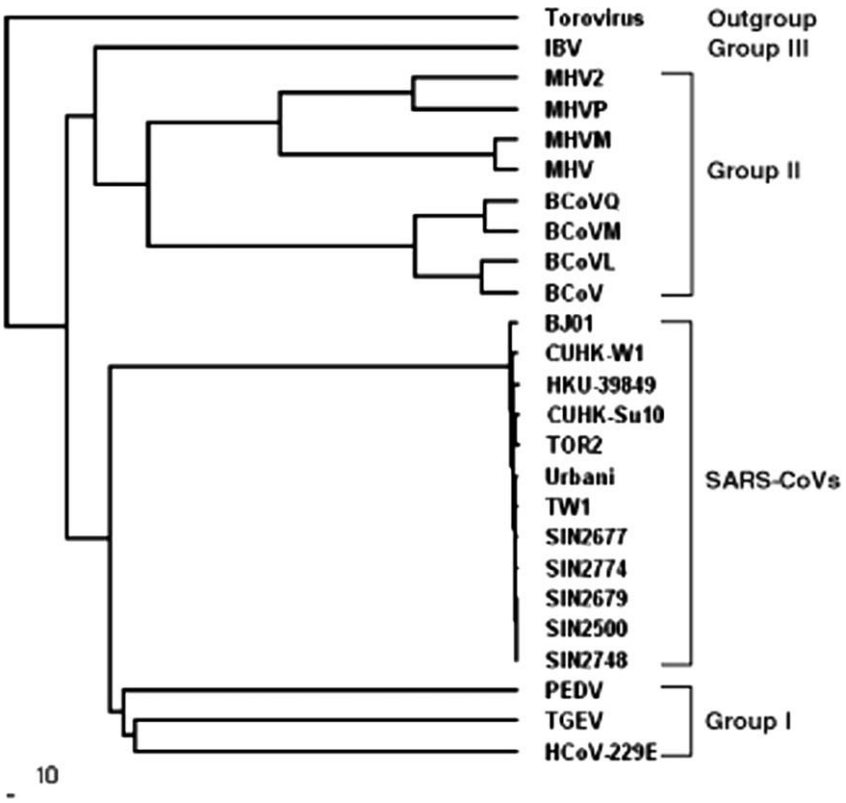

Fig. 3. Phylogenetic tree built by our distance d_{Exp} at $k = 6$, where background frequency of each word is estimated by the product of the corresponding nucleotide frequencies.

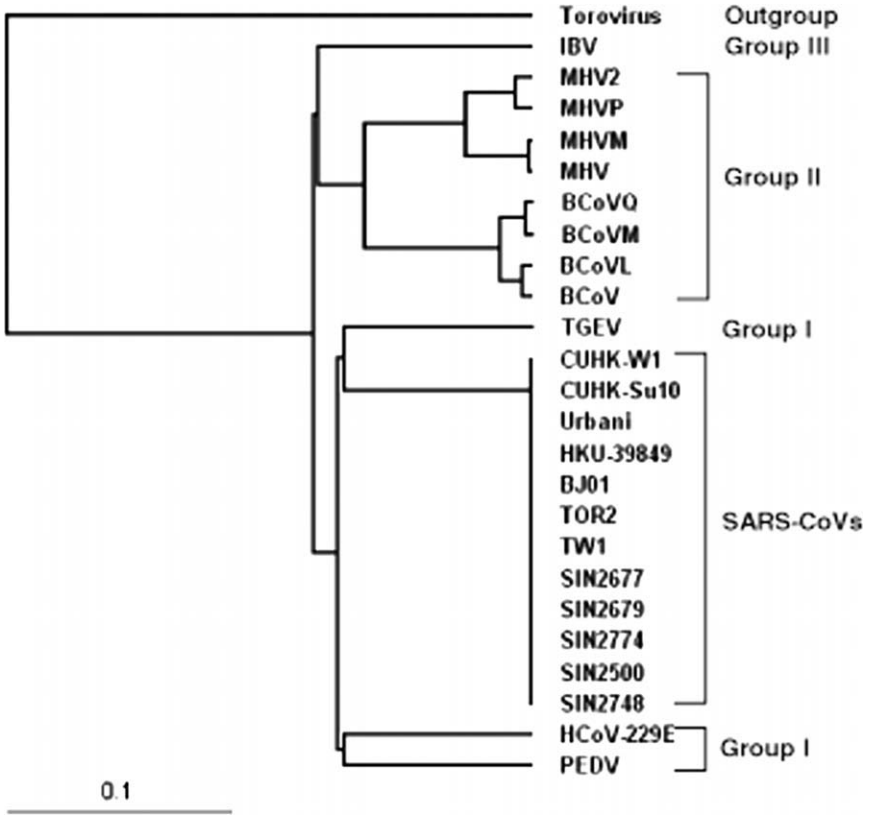

Fig. 4. Phylogenetic tree built by the Weighted Sequence Entropy (WSE) at word length $k = 6$.

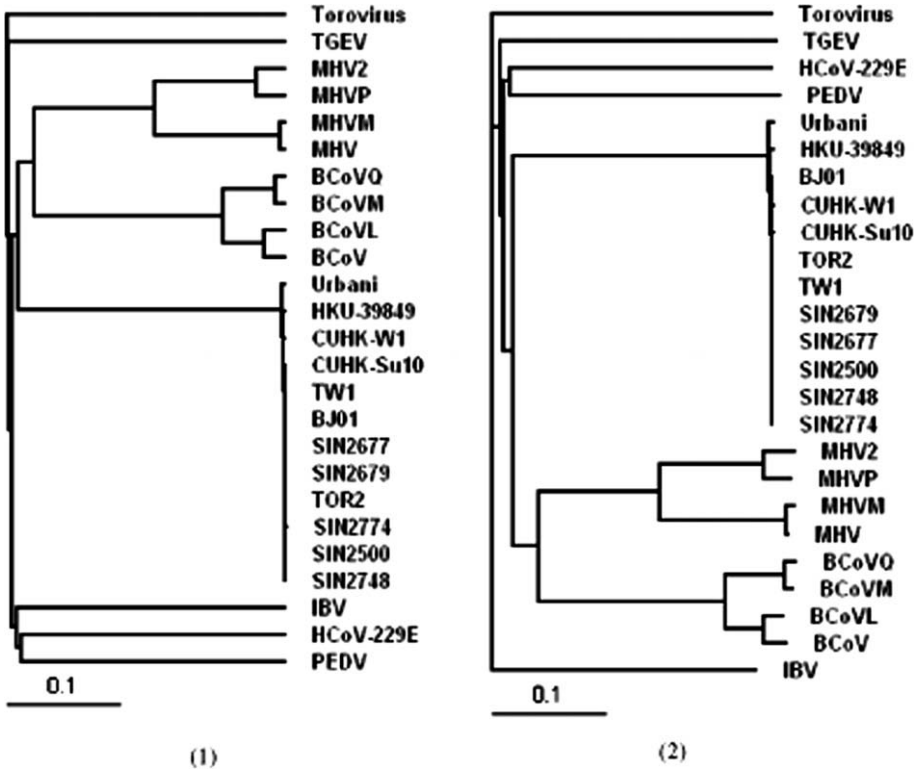

Fig. 5. (1) UPGMA and (2) Neighbor-Joining trees of 25 viruses built by the string Composition Vector approach.

topic in molecular genetics [33]. Different types of molecular data and analysis methods result in different trees. By the maximum likelihood method, some proteins support the Ferungulates (Primates, Rodents) grouping while other proteins support the Rodents (Ferungulates, Primates) grouping [34]. Whereas our result suggests an alternative topology of Primates (Ferungulates, Ro-

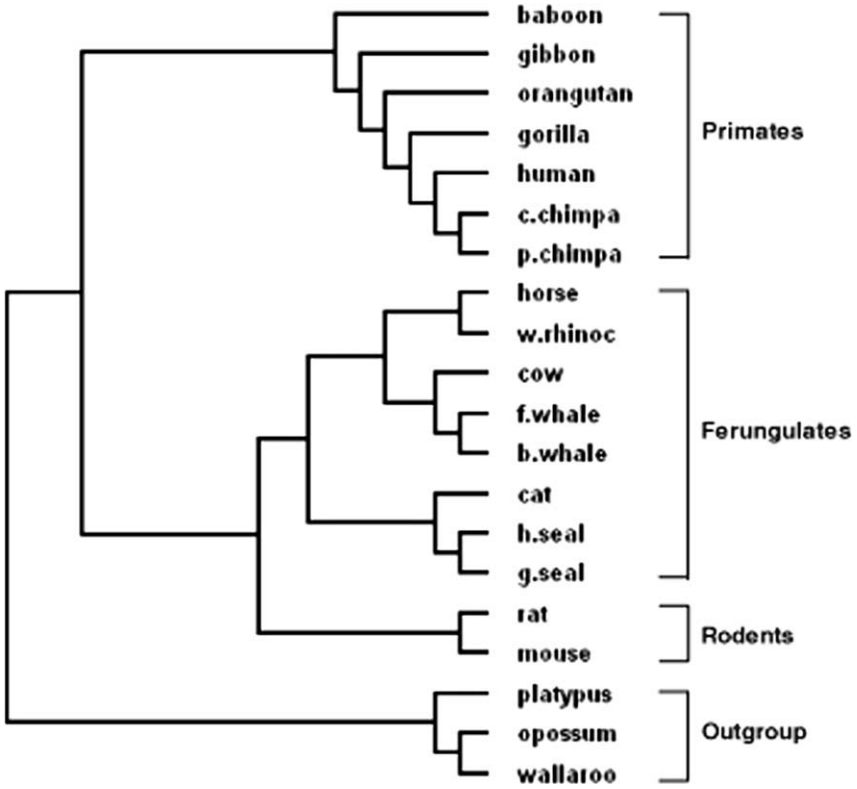

Fig. 6. The UPGMA tree built from the complete mtDNA sequences of 20 mammals. We use the distance metric d_{Exp}, and background frequencies of words are estimated by the Markov model of order $k-2$.

dents). In addition, we also applied some other word-based metrics mentioned above (the standard Euclidean distance, linear correlation coefficient and KL discrepancy) to the same dataset, but they did not give competitive results (not shown in this paper).

4. Conclusion and discussion

With the completion of many genome projects of Prokaryotes and Eukaryotes, genome level phylogeny constructions are available and expected to be more reliable compared to traditional experiments on only a single gene or a fragment of genome. However, multiple sequence alignment of genomic sequences is still a bottleneck, first due to the computational time, and second due to the inherent model assumptions. Therefore, there is a great need to develop new sequence comparisons free of these problems. In recent years, a quantity of alignment-free methods which are based on, e.g., k-words frequency [2], graphical representations [35–42], and information contents [32,43], have been proposed. Nevertheless, compared to alignment methods, these methods are still in the premature stage.

Sequence comparison based on the genomic composition of short words may be the most widely studied alignment-free method. It has relatively low computational complexity, and does not suffer greatly from genetic rearrangements and transposon activity, which serve as common ways of genome evolution. In most cases, biological sequences are represented as occurrence or frequency vectors in a high dimensional Euclidean space, and then the standard Euclidean distance, linear correlation coefficient, Kullback–Leibler (KL) discrepancy or cosine function between these vectors are calculated as measures of dissimilarity. In this paper, we investigate two word-based distance measurements in a probabilistic framework. Our hypothesis is that occurrence of a given word in a random DNA sequence follows the Poisson distribution. Then distance between two sequences is evaluated by the probability of generating one sequence under the Poisson model estimated from the other, or their different expression levels of words. In contrast to the traditional word-based distances, which use only frequencies of fixed-length words, our distances take background information of words (estimated by frequencies of some shorter words or the corresponding nucleotide composition) into account. In other words, our method has a potential to adjust the background information for distance measurements using composition vector. Through constructing phylogenetic trees of 25 viruses including SARS-CoVs and 20 Eutherian mammals, we find that our method gives a more competitive result compared to the ongoing word-based methods.

It is detected that each component CV_ω of the string Composition Vector is also a measure of expression in terms of word ω. In Eq. (7), the numerator $f_\omega - \bar{f}_\omega$ is the deviation of the observed frequency from the expected value, and denominator is introduced to eliminate the size effect. However, different from our measure (Eq. (3)), the value of CV_ω may be affected by those words with very low background frequency, i.e., when $\bar{f}_\omega$ is very small, the corresponding CV_ω will be very large. While our measure is free of this problem as it ranges from 0 to 1. In other words, our method can avoid the noise accompanied by words with exceptional background frequencies.

However, compared to those word-based measurements which consider only composition vectors, our distances have relatively high computational costs. For example, occurrences of many words are much higher than 60 in some bacterial genomes (when $k = 10$), which makes our Poisson-based distances computationally infeasible. So a reliable and efficient approximation of Poisson probability is critical to our method. In addition, the accuracy of our approach depends strongly on the Poisson model of word occurrences. This assumption is generally valid when the sequence length is suffi-

ciently large. But for words with overlapping structure, e.g., TATATA and CCGCCG, their occurrences in a random sequence may vary significantly from the Poisson distribution. While at the same time, experiments showed that these self-overlapping words are more prone to be functional patterns in regular regions of genomes. In the future study, we will explore some models to describe and compare these words.

Acknowledgment

This work was supported in part by Leading Academic Discipline Project of Shanghai Normal University (No. DZL803) and Shanghai Leading Academic Discipline Project (No. S30405).

References

[1] A. Roy, C. Raychaudhury, A. Nandy, Novel techniques of graphical representation and analysis of DNA sequences – a review, J. Biosci. 23 (1998) 55.
[2] S. Vinga, J.S. Almeida, Alignment-free sequence comparison – a review, Bioinformatics 19 (2003) 513.
[3] T.J. Wu, Y.C. Hsieh, L.A. Li, Statistical measures of DNA dissimilarity under Markov chain models of base composition, Biometrics 57 (2001) 441.
[4] D. Burstein, I. Ulitsky, T. Tuller, B. Chor, Information theoretic approaches to whole genome phylogenies, in: Proceedings of the Ninth Annual International Conference on Research in Computational Molecular Biology (RECOMB 2005), Cambridge, MA, 2005.
[5] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403.
[6] W.R. Pearson, D.J. Lipman, Improved tools for biological sequence comparison, Proc. Natl. Acad. Sci. USA 85 (1988) 2444.
[7] A. Campbell, J. Mrazek, S. Karlin, Genome signature comparisons among prokaryote plasmid, and mitochondrial DNA, Proc. Natl. Acad. Sci. USA 96 (1999) 9184.
[8] S. Karlin, I. Ladunga, Comparisons of eukaryotic genomic sequences, Proc. Natl. Acad. Sci. USA 91 (1994) 12832.
[9] S. Karlin, J. Mrázek, Compositional differences within and between eukaryotic genomes, Proc. Natl. Acad. Sci. USA 94 (1997) 10227.
[10] G. Reinert, S. Schbath, M.S. Waterman, Probabilistic and statistical properties of words: an overview, J. Comput. Biol. 7 (2000) 1.
[11] H.J. Jeffrey, Chaos game representation of gene structure, Nucleic Acids Res. 18 (1990) 2163.
[12] J. Shen, S. Zhang, H.C. Lee, B.L. Hao, SeeDNA: a visualization tool for *k*-string content of long DNA sequences and their randomized counterparts, Geno. Prot. Bioinfo. 2 (2004) 192.
[13] A.J. Gentles, S. Karlin, Genome-scale compositional comparisons in eukaryotes, Genome Res. 11 (2001) 540.
[14] H. Nakashima, M. Ota, K. Nishikawa, T. Ooi, Genes from nine genomes are separated into their organisms in the dinucleotide composition space, DNA Res. 5 (1998) 251.
[15] T.J. Wu, J.P. Burke, D.B. Davison, A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words, Biometrics 53 (1997) 1431.
[16] J. Qi, B. Wang, B.L. Hao, Whole proteome prokaryote phylogeny without sequence alignment: a *k*-string composition approach, J. Mol. Evol. 58 (2004) 1.
[17] M.M. Lai, D. Cavanagh, The molecular biology of coronaviruses, Adv. Virus Res. 48 (1997) 1.
[18] J.H. Strauss, E.G. Strauss, Viruses and Human Diseases, Academic Press, San Diego, 2002.
[19] M.M.C. Lai, K.V. Holmes, Coronaviridae: the viruses and their replication, in: D.M. Knipe, P.M. Howley (Eds.), Fields Virology, fourth ed., Lippincott–Williams & Wilkins, New York, 2001.
[20] M.A. Marra, S.J. Jones, C.R. Astell, et al., The genome sequence of the SARS-associated coronavirus, Science 300 (2003) 1399.
[21] P.A. Rota, M.S. Oberste, S.S. Monroe, et al., Characterization of a novel coronavirus associated with severe acute respiratory syndrome, Science 300 (2003) 1394.
[22] P. Liò, N. Goldman, Phylogenomics and bioinformatics of SARS-CoV, Trends Microbiol. 12 (2004) 106.
[23] A.C. Yang, A.L. Goldberger, C.K. Peng, Genomic classification using an information-based similarity index: application to the SARS coronavirus, J. Comput. Biol. 12 (2005) 1103.
[24] E.J. Snijder, M.C. Horzinek, Toroviruses: replication, evolution and comparison with other members of the coronavirus-like superfamily, J. Gen. Virol. 74 (1993) 2305.
[25] B.E. Blaisdell, A measure of the similarity of sets of sequences not requiring sequence alignment, Proc. Natl. Acad. Sci. USA 83 (1986) 5155.
[26] P. Petrilli, Classification of protein sequences by their dipeptide composition, Comput. Appl. Biosci. 9 (1993) 205.
[27] J. Qi, H. Luo, B. Hao, CVTree: a phylogenetic tree reconstruction tool based on whole genomes, Nucleic Acids Res. 32 (2004) 45.
[28] J. Wang, X. Zheng, WSE: a new sequence distance measure based on word frequencies, Math. Biosci. 215 (2008) 78.
[29] R.D. Page, TreeView: an application to display phylogenetic trees on personal computers, Comput. Appl. Biosci. 12 (1996) 357.
[30] W.C. Zheng, L.L. Chen, H.Y. Ou, F. Gao, C.T. Zhang, Coronavirus phylogeny based on a geometric approach, Mol. Phylogenet. Evol. 36 (2005) 224.
[31] T.G. Ksiazek, D. Erdman, C.S. Goldsmith, et al., A novel coronavirus associated with severe acute respiratory syndrome, N. Engl. J. Med. 348 (2003) 1953.
[32] H.H. Otu, K. Sayood, A new sequence distance measure for phylogenetic tree construction, Bioinformatics 19 (2003) 2122.
[33] A. Reyes, C. Gissi, G. Pesole, F.M. Catzeflis, C. Saccone, Where do rodents fit? Evidence from the complete mitochondrial genome of *Sciurus vulgaris*, Mol. Biol. Evol. 17 (2000) 979.
[34] Y. Cao, A. Janke, P.J. Waddell, M. Westerman, O. Takenaka, S. Murata, N. Okada, S. Paabo, M. Hasegawa, Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders, J. Mol. Evol. 47 (1998) 307.
[35] M. Gates, A simple way to look at DNA, J. Theor. Biol. 119 (1986) 319.
[36] B. Liao, Y.S. Liu, R.F. Li, W. Zhu, Coronavirus phylogeny based on triplets of nucleic acids bases, Chem. Phys. Lett. 421 (2006) 313.
[37] B. Liao, X.Y. Xiang, W. Zhu, Coronavirus phylogeny based on 2D graphical representation of DNA sequence, J. Comput. Chem. 27 (2006) 1196.
[38] A. Nandy, A new graphical representation and analysis of DNA sequence structure. I. Methodology and application to globin genes, Curr. Sci. 66 (1994) 309.
[39] A. Nandy, P. Nandy, On the uniqueness of quantitative DNA difference descriptors in 2D graphical representation models, Chem. Phys. Lett. 368 (2003) 102.
[40] M. Randić, A.T. Balaban, M. Novič, A. Založnik, T. Pisanski, A novel graphical representation of proteins, Period. Biol. 107 (2005) 403.
[41] M. Randić, D. Butina, J. Zupan, Novel 2-D graphical representation of proteins, Chem. Phys. Lett. 419 (2006) 528.
[42] C.T. Zhang, A symmetrical theory of DNA sequences and its application, J. Theor. Biol. 187 (1997) 297.
[43] M. Li, J.H. Badger, X. Chen, S. Kwong, P. Kearney, H.Y. Zhang, An information-based sequence distance and its application to whole mitochondrial genome phylogeny, Bioinformatics 17 (2001) 149.