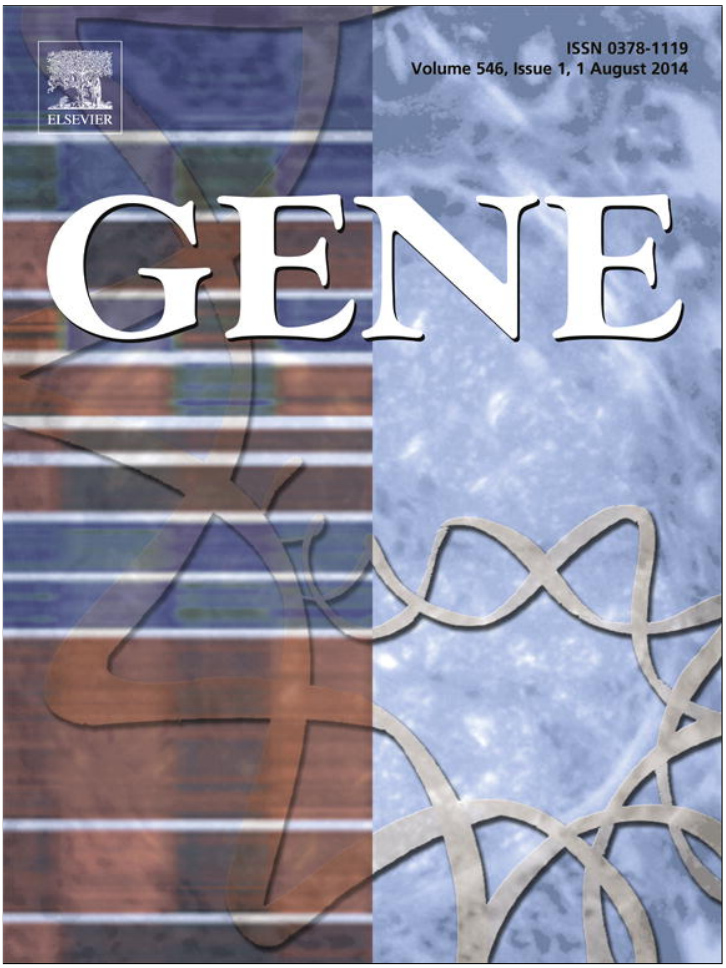


Gene 546 (2014) 25–34

Contents lists available at ScienceDirect

Gene

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

K-mer natural vector and its application to the phylogenetic analysis of genetic sequences

Jia Wen [a,b], Raymond H.F. Chan [b], Shek-Chung Yau [c], Rong L. He [d], Stephen S.T. Yau [e,*]

[a] School of Information Science, Suihua University, Suihua 152061, China
[b] Department of Mathematics, The Chinese University of Hong Kong, Shatin 999077, Hong Kong
[c] Information Technology Services Center, Hong Kong University of Science and Technology, Kowloon 999077, Hong Kong
[d] Department of Biological Sciences, Chicago State University, Chicago, IL 60628, USA
[e] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 8 February 2014
Received in revised form 4 May 2014
Accepted 20 May 2014
Available online 22 May 2014

Keywords:
K-mer model
Natural vector
Phylogenetic analysis

ABSTRACT

Based on the well-known *k*-mer model, we propose a *k*-mer natural vector model for representing a genetic sequence based on the numbers and distributions of *k*-mers in the sequence. We show that there exists a one-to-one correspondence between a genetic sequence and its associated *k*-mer natural vector. The *k*-mer natural vector method can be easily and quickly used to perform phylogenetic analysis of genetic sequences without requiring evolutionary models or human intervention. Whole or partial genomes can be handled more effective with our proposed method. It is applied to the phylogenetic analysis of genetic sequences, and the obtaining results fully demonstrate that the *k*-mer natural vector method is a very powerful tool for analysing and annotating genetic sequences and determining evolutionary relationships both in terms of accuracy and efficiency.

1. Introduction

Phylogenetic analysis of genetic sequences has become essential for researching the evolutionary relationships between all types of organisms (from bacteria to humans) (Nei, 1996). Phylogenetic analysis is also important for clarifying the evolutionary pattern of multigene families (Atchley et al., 1994; Goodwin et al., 1996; Ota and Nei, 1994), as well as for understanding adaptive evolution at the molecular level (Chandrasekharan et al., 1996; Jermann et al., 1995; Wistow, 1993). It also provides deep insight into the mechanism for the maintenance of polymorphic alleles in populations (Figueroa et al., 1988; Takahata, 1993). The results of phylogenetic analysis are represented by phylogenetic tree, in which sequences are grouped based on sequence similarities.

Methods for phylogenetic analysis commonly depend on multiple sequence alignment, which assumes some sort of evolutionary model, and yields results that are often controversial. Although most alignment-based methods can precisely represent evolutionary relationships between genetic sequences, they frequently lead to very complicated computation. Alignment-free methods, which are based on numerical characterizations of genetic sequences, are proposed to compensate for the ineffectiveness of traditional alignment-based methods.

Among all alignment-free methods, the *k*-mer model method may be the best developed one. The classic string representation based on the *k*-mer model was first used for the comparison of genome sequences by Blaisdell (1986), and the counts of *k*-mers appearing in the sequence were used for the comparison of regulatory sequences by Kantorovitz et al. (2007). Later, various frequency-based methods were introduced for sequence comparison presented by Wu et al. (1997, 2001, 2005), Korf and Rose (2009), Sims et al. (2009a, 2009b) and Jun et al. (2010). The advantage of *k*-mer model approach is that the phylogenetic tree can be constructed much faster than using sequence alignment, and it can be used for comparison of whole genomes. However, the deficiency of the *k*-mer model is that the relationships between the *k*-mers within a sequence are more or less neglected (Yang and Wang, 2013; Yu, 2013).

The original natural vector approach is an alternative alignment-free method which produces a one-to-one association between genetic sequences and vectors in a finite dimensional space (Deng et al., 2011). One of the strengths of this approach is that the natural vector incorporates the normalized central moments to account for the interrelationships between different portions of genetic sequences. But the obtaining results show that the original natural vector approach cannot accurately depict evolutionary relationships of species considered in phylogenetic analysis of genetic sequences.

In this paper, we integrate original natural vector with *k*-mer model to produce *k*-mer natural vector that contains both types of information: the information stored in the *k*-mer counts as well as information about the relationships between *k*-mers appearing in the sequence. We

Abbreviations: A, adenosine; C, cytidine; G, guanosine; T, thymidine; bp, base pairs; UPGMA, Unweighted Pair Group Method with Arithmetic Mean; NJ, Neighbour Joining.

* Corresponding author.
E-mail address: yau@uic.edu (S.S.T. Yau).

http://dx.doi.org/10.1016/j.gene.2014.05.043
0378-1119/

can prove that the correspondence between a genetic sequence and its associated *k*-mer natural vector is one-to-one by mathematical proof. Moreover, the *k*-mer natural vector method is applied to the phylogenetic analysis of genetic sequences, and the obtaining results show that our new method can not only effectively overcome the deficient of former *k*-mer model methods, but also further improve accuracy in depicting evolutionary relationships of genetic sequences compared with sequence alignment methods and some published methods.

2. Materials and methods

2.1. K-mer model of genetic sequence

The *k*-mer model of a genetic sequence can be described as follows: Consider a genetic sequence *s* of length *L*, '$N_1N_2\ldots N_L$', where $N_l \in \{A,C,G,T\}, l = 1,2,\ldots,L$. A string of consecutive *k* nucleotides within a genetic sequence is called a *k*-mer. The *k*-mers appearing in a sequence can be

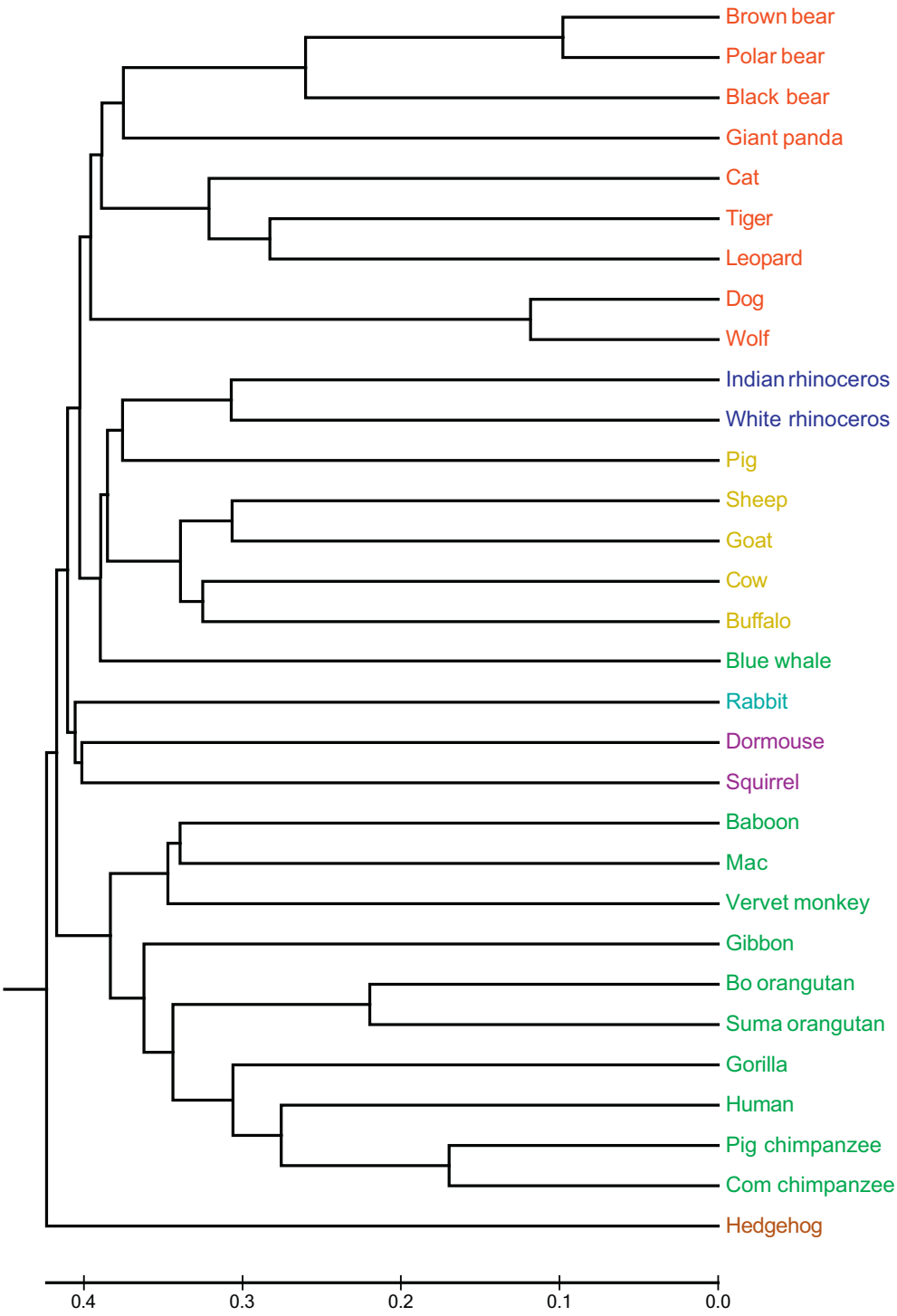


Fig. 1. Phylogenetic tree of 31 mitochondrial genome sequences based on 9-mer natural vector. All 31 genomes are correctly clustered into eight known clusters: Carnivora (red), Perissodactyla (blue), Artiodactyla (yellow), Cetacea (light green), Lagomorpha (light blue), Rodentia (purple), Primates (green) and Erinaceomorpha (light green), which agrees with results from standard biological taxonomy and evolutionary relationships of species.

enumerated by using a sliding window of length k, shifting one base each time from position 1 to $L - k + 1$, until the entire sequence has been scanned.

Given any k, there are 4^k different possible permutations of k-mers that may appear: [1], [2], ..., [4^k]. For any genetic sequence s, the k-mer counting vector $n^{(s,k)}$ is defined by $n^{(s,k)} = \left(n_{s[1]}, n_{s[2]}, \ldots, n_{s[4^k]}\right)$, where $n_{s[i]}$ is the number of times the k-mer [i] occurs in sequence s.

2.2. K-mer natural vector

The k-mer natural vector is defined to be the concatenation of the following three vectors, each of which is of length 4^k:

The k-mer counting vector $n^{(s,k)}$ as defined above.
The k-mer mean distance vector $\left(\mu_{[1]}, \mu_{[2]}, \ldots, \mu_{[4^k]}\right)$, where $\mu_{[i]}$ is defined to be the arithmetic mean of the distances from various

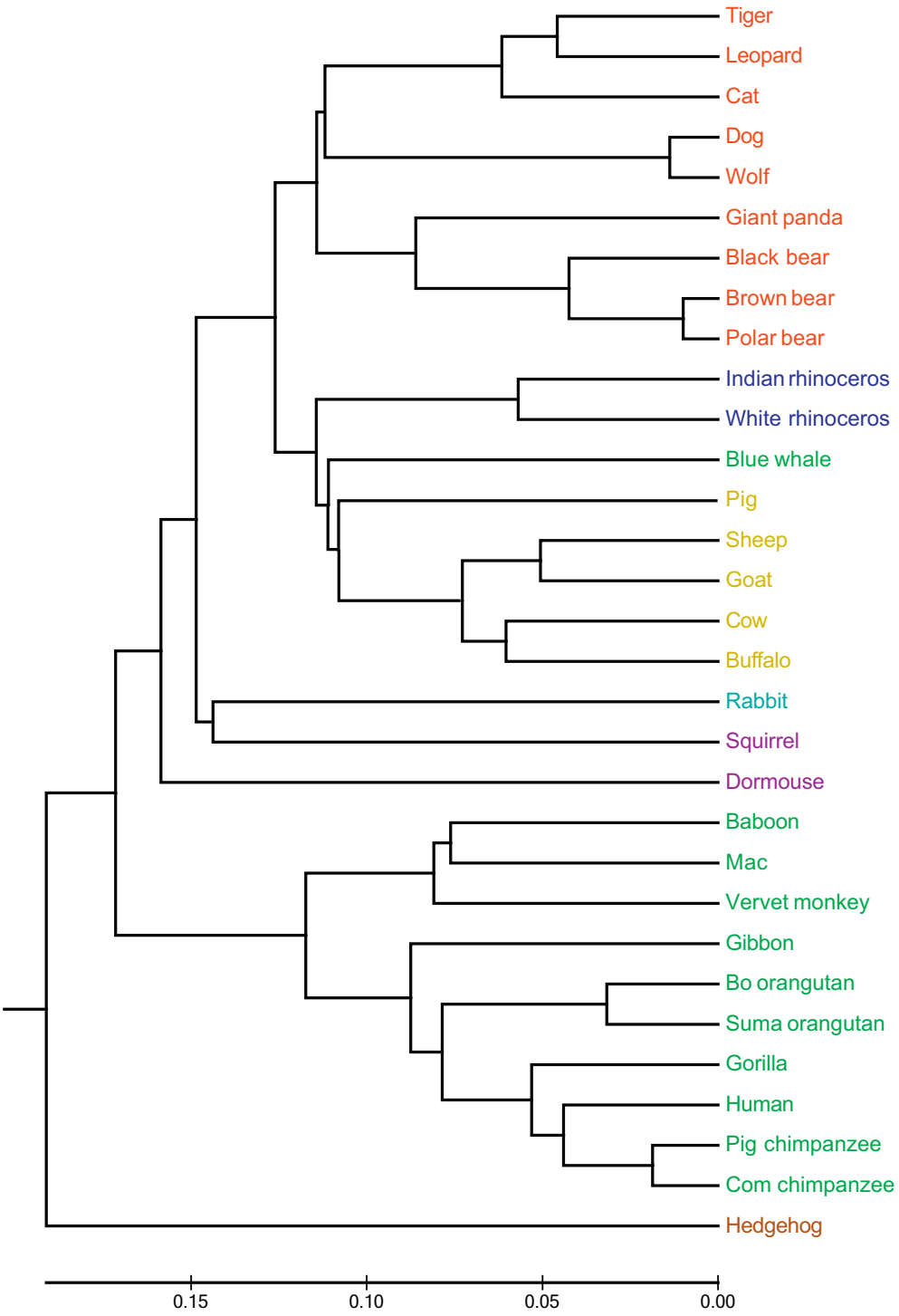
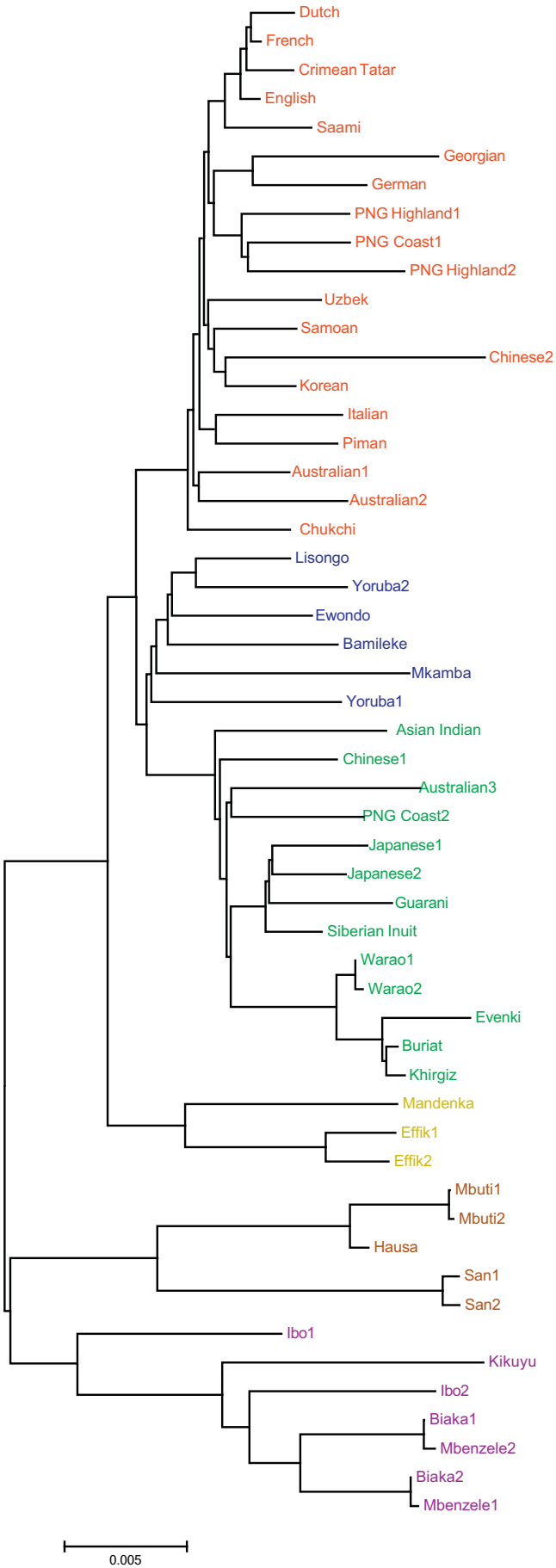


Fig. 2. Phylogenetic tree of 31 mitochondrial genome sequences obtained by multiple sequence alignment (clustalW).



occurrences of the k-mer $[i]$ to the first base in the sequence. If a specific k-mer $[i]$ does not exist in a genetic sequence, $\mu_{[i]}$ is defined to be zero.

The normalized central moment vector $\left(D_2^{[1]}, D_2^{[2]}, \ldots, D_2^{[4^k]}\right)$. In general, for any m, the normalized central moments are defined as follows:

$$D_m^{[i]} = \sum_{j=1}^{n_{[i]}} \frac{\left(s[i][j]-\mu_{[i]}\right)^m}{n_{[i]}^{m-1}(L-k+1)^{m-1}}, m = 1, 2, \ldots, n_{[i]},$$

where $n_{[i]}$ denotes the number of times $[i]$ appearing in the genetic sequence, L is the length of genetic sequence, $s[i][j]$ is the distance from the first base to the j-th $[i]$ in sequence s, and $\mu_{[i]}$ is the mean of distances from the various occurrences of $[i]$ to the first base. Thus, we get a sequence of normalized central moments which are natural parameters associated with k-mer distributions within the genetic sequence.

When $k = 1$, the k-mer natural vector is the same to the original natural vector. Thus the k-mer natural vector method is a generalization of the original natural vector model.

If the distribution of each k-mer is different, two genetic sequences cannot be similar even though they contain the same set of k-mers and the same total distance measurement. Although each subset of numerical parameters maybe not sufficient to annotate genetic sequences, the combined numerical parameters are sufficient to characterize each genetic sequence. We can mathematically prove that the correspondence between a genetic sequence and its associated k-mer natural vector is one-to-one for each given k in Text S1 of Appendix A. Because all the first central moments are zero, we do not need to include them as part of k-mer natural vector.

The k-mer natural vector is obtained by concatenating the first group of parameters (the frequency of occurrence of each k-mer in the sequence) and the second group of parameters (the mean distance of each k-mer to the first base) to the normalized central moments, and the k-mer natural vector implies the information on the relationships of k-mers for each fixed k. Because of this, our k-mer natural vector model overcomes the deficiency of previous k-mer model methods.

It is shown that the 3×4^k-dimensional vector $(n_{[i]}, \mu_{[i]}, D_2^{[i]})$ is enough to represent a genetic sequence, and not necessary to include normalized central moments higher than second order for the comparison of genetic sequences, in that, the high central moments hardly make any contribution, and the 3×4^k-dimensional natural vector mapping restricted on all the datasets is still one-to-one mapping.

For each fixed k, there are 4^k different possible k-mers in the sequence. The computational complexity of our k-mer natural vector is $o(n \cdot m^2 \cdot 4^k)$, where n is the maximum length of the sequences, and m is the number of sequences in the dataset. Our proposed method is fast, because it only needs to read the sequence once to compute k-mer natural vector. Moreover, the running time comparisons for our k-mer natural vector methods, clustalW, and Muscle are presented in Text S2 of Appendix A.

2.3. The choice of k

Because the parameter k has a great influence on the results of evolutionary relationships and on the complexity of computation for k-mer

Fig. 3. Phylogenetic tree of 53 human mitochondrial genome sequences based on 8-mer natural vector. The 53 mtDNAs are mainly divided into two parts: non-Africans (red and green) and Africans (blue, yellow, brown and purple), and humans in each group correctly cluster, which is consistent with known evidences of human evolution and human migration.

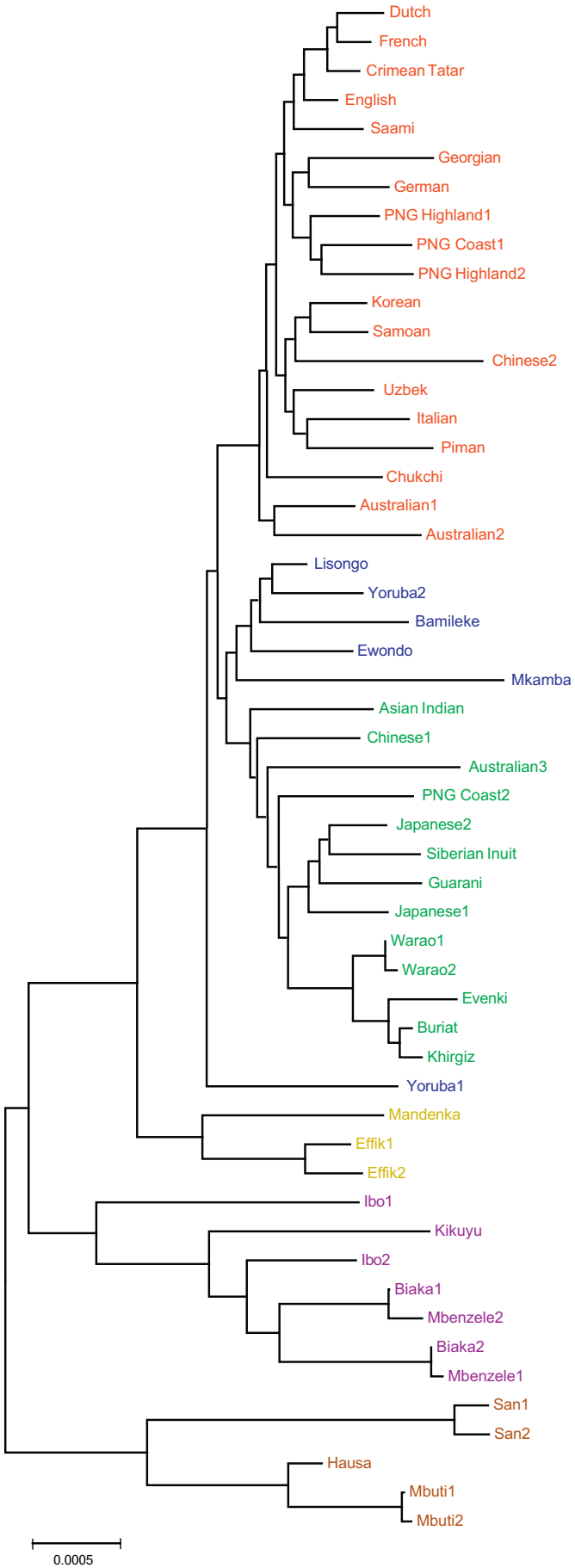


Fig. 4. Phylogenetic tree of 53 human mitochondrial genome sequences obtained by multiple sequence alignment (clustalW).

model, it is very important and difficult to choose a suitable k for different lengths of genetic sequences considered in phylogenetic analysis. Some researchers have explored the selection of the optimum value k^* for k-mer model. For example, Wu et al. proposed an optimal word size for dissimilarity measurement that depends on the length of sequences being considered, i.e., k^* should be increased when the sequence length increases (Wu et al., 2005). Another investigation was done by Sims et al. (2009a, 2009b), who reported that the optimal length of word lies within an approximate range with lower bound $\log_4 n$, where n is the length of sequence, and the upper bound given by the criterion that phylogenetic tree topology for length k must be parallel to that of $k+1$.

Searching for the optimum value k^* for k-mer model, we apply our proposed method to some real datasets (Chan et al., 2012; Deng et al., 2011; Huang et al., 2011; Ingman et al., 2000; Yu, 2013), and the optimal k^* over the range of k considered for k-mer natural vector model is chosen based on the following strategy: if the result of phylogenetic tree for value k is relatively stable to that of $k+1$, we choose $k^* = k$; otherwise k^* is equal to the maximum over the range of k values considered. We infer that the optimal k^* for our k-mer natural vector is within a range

$$[ceil(\log_4 \min(L)), ceil(\log_4 \max(L)) + 1],$$

where L is the set of lengths of genetic sequences considered in phylogenetic analysis. This explicit range for choosing the optimum value k^* is much shorter than that considered by previous k-mer model methods. Additionally, the optimal k^* obtained by k-mer natural vector is less than those selected by other k-mer model methods (Chan et al., 2012; Qi et al., 2004; Yu et al., 2005) for the same candidate dataset (18S rRNA dataset), which indicates that our k-mer natural vector method needs lower computational time, and can more easily extract the features that are hidden in genetic sequence.

2.4. Distance metric

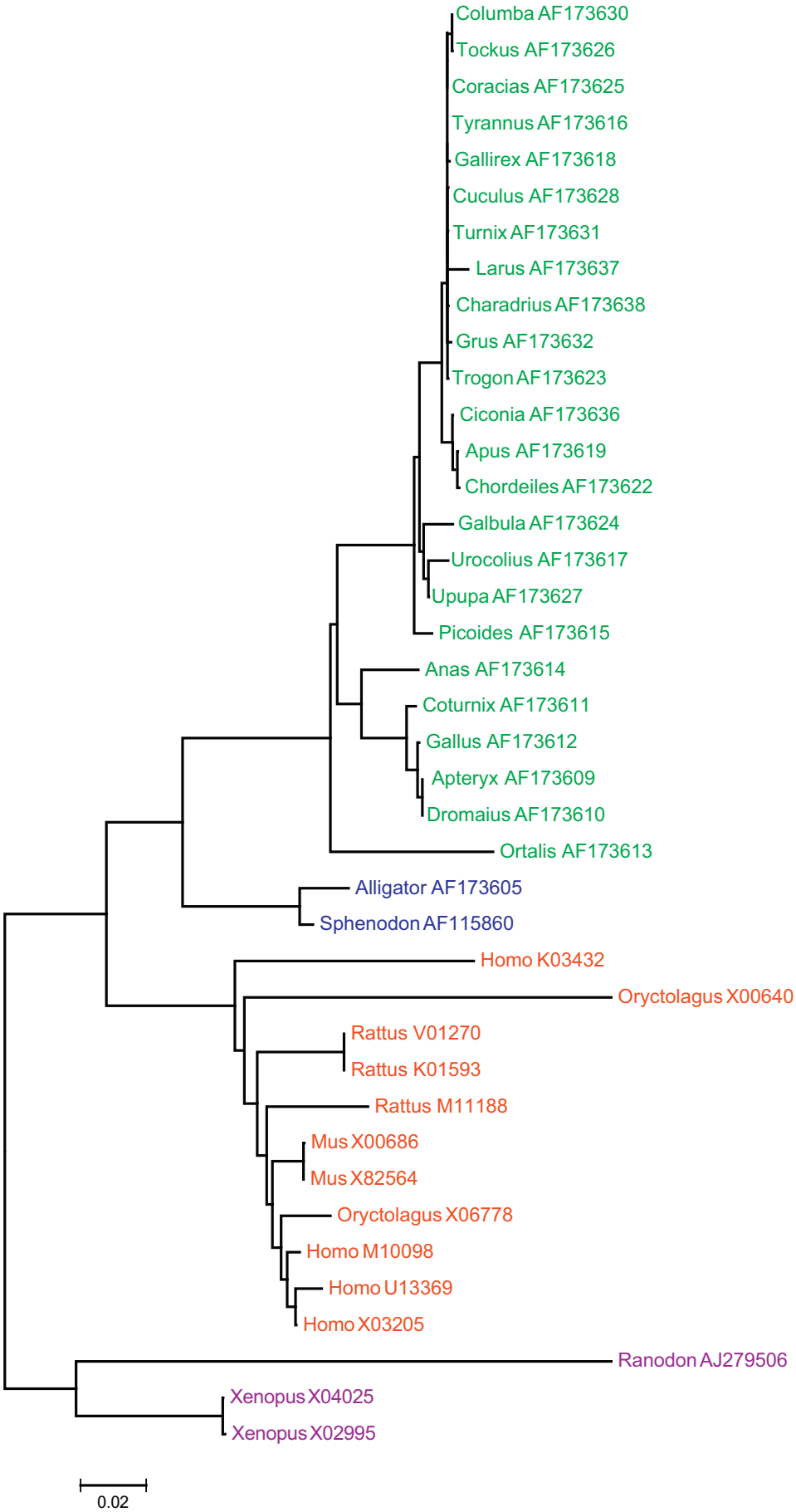
Since each genetic sequence can be uniquely represented by a k-mer natural vector, a distance metric can be used to quantify the evolutionary relationships of genetic sequences. The similarity between a pair of genetic sequences can be computed by the correlation angle between their natural vectors, because the correlation angle can eliminate the effects of high dimensionality (Berry et al., 1999; Wen and Zhang, 2009). In this paper, we select the distance metric defined below to measure the similarities of genetic sequences, which has been widely used in the k-mer model (Qi et al., 2004; Stuart and Berry, 2004; Stuart et al., 2002).

Let v_1 and v_2 be the k-mer natural vectors of genetic sequences s_1 and s_2, respectively, the distance between sequences s_1 and s_2 can be computed as follows:

$$d(s_1, s_2) = 1 - \cos(v_1, v_2) = 1 - \frac{v_1 \cdot v_2}{|v_1||v_2|},$$

where $\cos(v_1,v_2)$ is the cosine angle of vectors v_1 and v_2, and $|v_1|$,$|v_2|$ are the norms of vectors v_1 and v_2, respectively.

Once the distance matrix constructed by the distances among all genetic sequences considered for phylogenetic analysis is obtained, the evolutionary tree can be drawn by the methods of Unweighted Pair Group Method with Arithmetic Mean (UPGMA) or Neighbour Joining (NJ) using MEGA 5.10 (Tamura et al., 2011).



3. Results and discussion

To demonstrate the validity of *k*-mer natural vector method, we apply our proposed method to the phylogenetic analysis of real datasets: the mitochondrial genome sequences and 18S rRNA sequences in which both long and short genetic sequences are considered. All genetic sequences are treated as linear sequences.

3.1. Phylogenetic analysis of 31 mammal mitochondrial genomes

We first analyse the mitochondrial genome sequences of 31 species using our proposed method. This data was previously analysed using the original natural vector approach (Deng et al., 2011). The descriptions of the 31 mitochondrial genome sequences are listed in Table S1 of Appendix A, the lengths of which are from 16,338 to 17,447 base pairs (bp). The mitochondrial genetic sequences that are not highly conserved have a rapid mutation rate, so they are suitable for exploring the evolutionary relationships of different species (Huang et al., 2011; Yu et al., 2010). The phylogenetic tree of 31 mitochondrial genomes is shown in Fig. 1 by UPGMA method when $k = 9$.

Looking at Fig. 1, all 31 genomes are correctly clustered into eight known clusters: Carnivora (red), Perissodactyla (blue), Artiodactyla (yellow), Cetacea (light green), Lagomorpha (light blue), Rodentia (purple), Primates (green) and Erinaceomorpha (brown). Since whales evolved from the primitive artiodactyl, blue whale clusters with artiodactyls to form Cetartiodactyla, which integrate with rhinoceroses to constitute Euungulata. Hence, our results can be considered as the evidence for Euungulata Theory. Additionally, rabbit clusters with dormouse and squirrel, in that, they are all in Glires. The resulting phylogenetic tree agrees with those from standard biological taxonomy, evolutionary relationships of species and some published papers (Huang et al., 2011; Kullberg et al., 2006; Liu et al., 2001; Raina et al., 2005; Yu et al., 2010). Compared with Fig. 3 of (Deng et al., 2011) drawn by the original natural vector method, the accuracy of evolutionary relationships has been greatly improved, which can be easily seen from the evolutionary relationships within the subgroups of Primates and Carnivora, respectively.

To further show the utility of our *k*-mer natural vector method, we perform multiple sequence alignment on the same dataset that we considered, using MEGA 5.10 implementation of the clustalW algorithm. The phylogenetic tree drawn for multiple sequence alignment is shown in Fig. 2 by UPGMA method, where the species are coloured the same as Fig. 1. Here, we only consider the differences between phylogenetic trees corresponding to the *k*-mer natural vector and clustalW, respectively. When clustalW is applied to 31 mitochondrial genome sequences, squirrel seems closer to rabbit in Fig. 2, rather than dormouse in Fig. 1, which does not agree with standard biological taxonomy, in that, squirrel and dormouse are rodents.

3.2. Phylogenetic analysis of 53 human mitochondrial genomes

We also apply our method to investigate variations in human mitochondrial genomes and to explore the origin of modern humans. Because mtDNA has a high substitution rate (Brown et al., 1979), less recombination (Olivio et al., 1983), and maternal inheritance (Giles et al., 1980), it is usually utilized as a tool in human evolution. Due to the variations of substitution rates and parallel mutation, these studies focusing on the control region of mtDNA might lead to incorrect phylogenetic inferences (Maddison et al., 1992; Tamura and Nei, 1993).

To improve the information obtained from mtDNA for studies of human evolution, Ingman et al. described the global mtDNA diversity in humans based on sequence alignment of complete mtDNA sequences (excluding D-loops) from 53 diverse origins (Ingman et al., 2000). It has been verified that the portion of a mtDNA sequence that is outside any D-loops evolves in a roughly 'clock-like' manner, enabling a more accurate measure of mutation rate, and therefore improved time estimates for evolutionary events. The 53 human mtDNAs (excluding D-loops) are unique and vary in length from 15,440 to 15,450 base pairs (bp). They are described in Table S2 of Appendix A and the phylogenetic tree for them is shown in Fig. 3 by NJ method when $k = 8$.

From Fig. 3, the 53 mtDNA sequences are divided into two parts: non-Africans (red and green) and Africans (blue, yellow, brown, and purple). Humans in each group correctly cluster, which is consistent with known evidences of human evolution and human migration. Compared with Fig. 2 of Ingman et al. (2000), the evolutionary relationships between all Africans and most non-Africans are the same, and differences only exist in several non-Africans.

Moreover, we also apply the clustalW to these 53 human mtDNA sequences, and the obtaining phylogenetic tree is shown in Fig. 4 by NJ method, which is the similar to the results of our proposed method shown in Fig. 3. Moreover, our *k*-mer natural vector method seems to get better results.

For example, sequence alignment method would imply that two mtDNA samples from Japanese were not closely connected, but our method (see Fig. 3) shows the contrary. If we take Guarani and Siberian-Inuit as references, the lengths of four mtDNAs considered are all 15,449. The mismatches between Japanese1 and Japanese2, Guarani, and Siberian-Inuit are 12, 16, and 15, respectively, and mismatches between Japanese2 and Guarani and Siberian-Inuit equal 14 and 13, respectively. Hence, the two Japanese should closely connect in the phylogenetic tree, and phylogenetic tree obtained by our method looks more reasonable.

3.3. Phylogenetic analysis of 40 tetrapod 18S rRNA sequences

Additionally, our method is used to analyse the phylogeny of 40 tetrapod 18S rRNA sequences. The 18S sequence was considered odd, providing significantly different estimates of phylogeny in higher organisms (Huelsenbeck et al., 1996). The phylogenetic relationship among tetrapod species has been widely discussed in the area of phylogeny and evolution. A controversial problem among tetrapod is whether birds are more closely related to crocodilians, or to mammals. The evolutionary analysis of tetrapod 18S rRNAs generates a clustering of birds with mammals (Xia et al., 2003), whereas evidences from molecules, palaeontology, and morphology showed that birds should cluster with crocodilians (Hedges et al., 1990), which is more acceptable to biologists. We investigate this question by applying our method to the tetrapod dataset shown in Fig. 3 of (Chan et al., 2012) which contains 40 sequences whose lengths are from 1733 to 2235 base pairs (bp).

The phylogenetic tree based on our proposed method is shown in Fig. 5 by NJ method when $k = 6$. This phylogenetic tree contains four clades: Birds (green), Crocodilians (blue), Mammals (red) and Amphibians (purple), and the species in each clade are correctly grouped together. The results are similar to those obtained from sequence alignment and what is found in some phylogenetic analyses (Ausio et al., 1999; Caspers et al., 1996; Chan et al., 2012; Dixon and Hillis, 1993; Hedges, 1994; Hedges et al., 1990; Janke and Arnason, 1997; Rzhetsky and Nei, 1992; Seutin et al., 1994; Xia et al., 2003; Zardoya and Meyer, 1998). It can be seen that birds group with crocodilians rather than group with mammals in Fig. 5. This result conforms to results from traditional classification and the results in Hedges et al. (1990) and Chan et al. (2012). Compared with Fig. 3 of Chan et al. (2012), our result is relatively better. Rattus and Mus group together, and Homo is closer to

Fig. 5. Phylogenetic tree of 40 18S rRNA sequences based on 6-mer natural vector. The phylogenetic tree of 18S rRNAs contains four clades: Birds (green), Crocodilians (blue), Mammals (red) and Amphibians (purple), and the species in each clade correctly group together that conform to results from traditional classification.

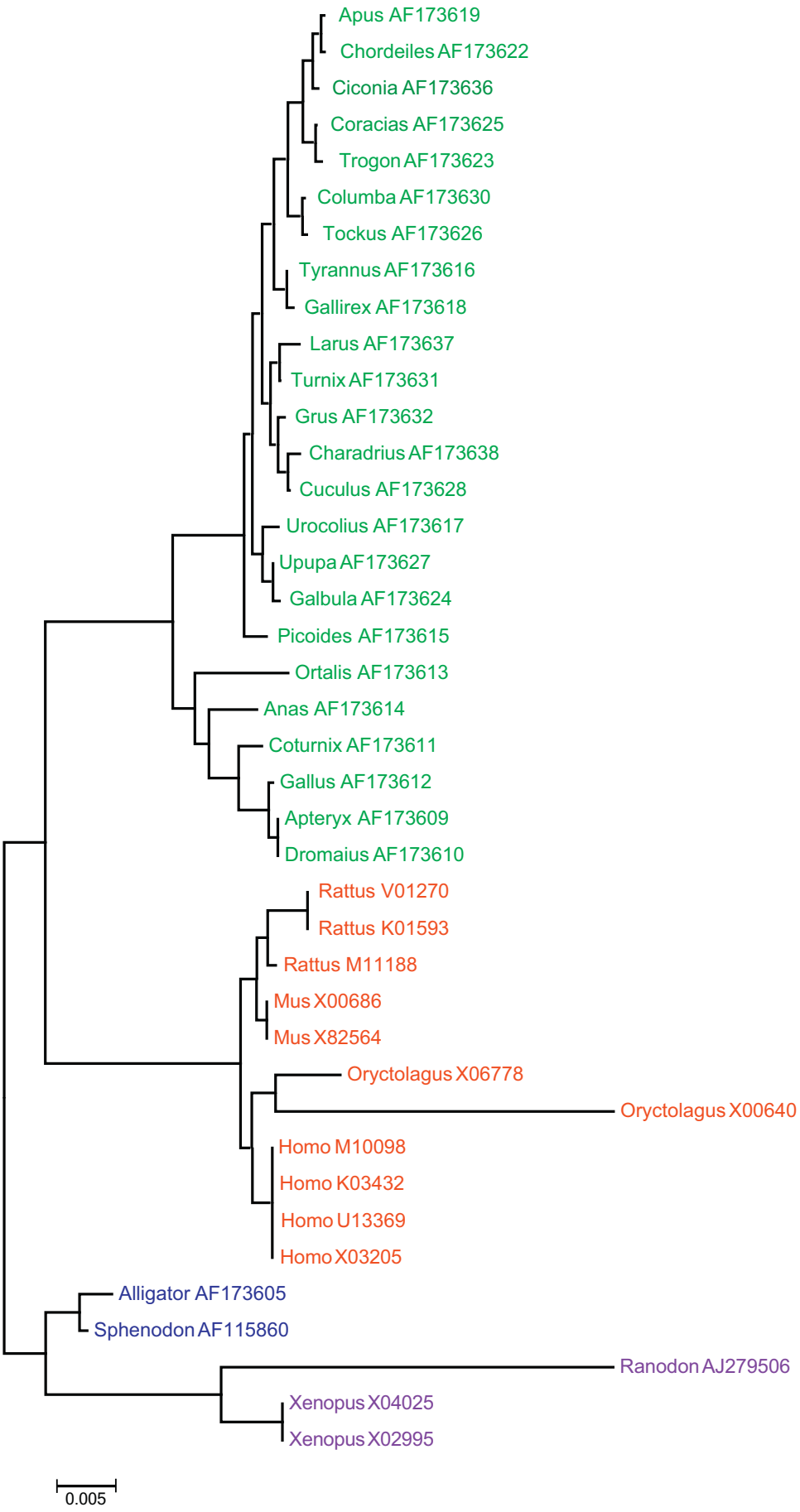


Fig. 6. Phylogenetic tree of 40 18S rRNA sequences obtained by multiple sequence alignment (clustalW).

Oryctolagus, rather than Mus and Rattus, which conform to the evolutionary relationships of species and results obtained by sequence alignment. Although in our Fig. 5, Homo K03432 is not clustered to the rest of Homo by NJ algorithm, however after inspecting the distance matrix, we find that the nearest neighbour of Homo K03432 is Homo M10089. Similarly, the nearest neighbour of Oryctolagus X00640 is Oryctolagus X06776.

Finally, we applied clustalW to the tetrapod 18S rRNA sequences, and the result is shown in Fig. 6. Although Homo and Oryctolagus do not group well, our proposed method has yielded a valuable result: birds group with crocodilians in Fig. 5, rather than mammals shown in Fig. 6, which conforms to traditional classification and evidences from molecules, morphology, and palaeontology. It is important to certify that bird should group with crocodilians, rather than with mammals, which would be more meaningful in biological evolution.

4. Conclusions

In this paper, the *k*-mer natural vector method is proposed by combining the original natural vector with the *k*-mer model for genetic sequences. The number and distribution of *k*-mers in a genetic sequence are the components of *k*-mer natural vector, which contains information of relationships between *k*-mers in a sequence. The correspondence between a genetic sequence and its associated *k*-mer natural vector can be mathematically proven to be one-to-one. With this representation, each genetic sequence can be characterized by a multidimensional vector. Our proposed method makes it easy to compare genetic sequences, which is more effective for handling whole or partial genomes than sequence alignment methods. The phylogenetic analysis of genetic sequences done by our proposed method does not assume some sort of evolutionary model, and avoids high computational complexity associated with sequence alignment. Its applications to real datasets have shown that the *k*-mer natural vector method is a powerful tool for the phylogenetic analysis of genetic sequences. It not only improves the accuracy of evolutionary relationships to some extent, but it also reduces computational time for phylogenetic analysis. However, the *k*-mer natural vector method is still in the process of being improved.

Conflict of interest

We certify that there is no conflict of interest. There is no limitation on access to data or other material critical to the work being reported.

Acknowledgements

We thank Dr. Max Benson for critically reading and editing our manuscript. This work is supported by Youth Funding of Suihua University (KQ1202004, KQ1202002), Scientific Research Funding of Heilongjiang Education Department (12513097), U.S. NSF grant (DMS-1120824, 1119612), NIH grant (5 SC3 GM098180-04), China NSF grant (31271408), Tsinghua University start up funding, and Tsinghua University independent research project grant.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2014.05.043.

References

Atchley, W.R., Fitch, W.M., Bronner, F.M., 1994. Molecular evolution of the MyoD family of transcription factors. Proc. Natl. Acad. Sci. U. S. A. 91, 11522–11526.
Ausio, J., Soley, J.T., Burger, W., Lewis, J.D., Barreda, D., Cheng, K.M., 1999. The histidine-rich protamine from ostrich and tinamou sperm: a link between reptile and bird protamines. Biochemistry 38, 180–184.
Berry, M.W., Drmac, Z., Jessup, E.R., 1999. Matrices, vector spaces, and information retrieval. SIAM Rev. 41, 335–362.
Blaisdell, B.E., 1986. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc. Natl. Acad. Sci. U. S. A. 83, 5155–5159.
Brown, W.M., George, M.J., Wilson, A.C., 1979. Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci. U. S. A. 76, 1967–1971.
Caspers, G.J., Reinders, G.J., Leunissen, J.A., Wattel, J., Dejong, W.W., 1996. Protein sequences indicate that turtles branched off from the amniote tree after mammals. J. Mol. Evol. 42, 580–586.
Chan, R.H., Chan, T.H., Yeung, H.M., Wang, R.W., 2012. Composition vector method based on maximum entropy principle for sequence comparison. IEEE/ACM Trans. Comput. Biol. Bioinform. 9, 79–87.
Chandrasekharan, U.M., Sanker, S., Glynias, M.J., Karnik, S.S., Husain, A., 1996. Angiotensin II-forming activity in a reconstructed ancestral chymase. Science 271, 502–505.
Deng, M., Yu, C., Liang, Q., He, R.L., Yau, S.S., 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. PLoS One 6, e17293.
Dixon, M.T., Hillis, D.M., 1993. Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. Mol. Biol. Evol. 10, 256–267.
Figueroa, F., Gunther, E., Klein, J., 1988. MHC polymorphism pre-dating speciation. Nature 335, 265–267.
Giles, R.E., Blanc, H., Cann, H.M., Wallace, D.C., 1980. Maternal inheritance of human mitochondrial DNA. Proc. Natl. Acad. Sci. U. S. A. 77, 6715–6719.
Goodwin, R.L., Baumann, H., Berger, F.G., 1996. Patterns of divergence during evolution of α_1-proteinase inhibitors in mammals. Mol. Biol. Evol. 13, 346–358.
Hedges, S.B., 1994. Molecular evidence for the origin of birds. Proc. Natl. Acad. Sci. U. S. A. 91, 2621–2624.
Hedges, S.B., Moberg, K.D., Maxson, L.R., 1990. Tetrapod phylogeny inferred from 18S and 28S ribosomal RNA sequence and a review of the evidence for amniote relationships. Mol. Biol. Evol. 7, 607–633.
Huang, G., Zhou, H., Li, Y.F., Xu, L., 2011. Alignment-free comparison of genome sequences by a new numerical characterization. J. Theor. Biol. 281, 107–112.
Huelsenbeck, J.P., Bull, J.J., Cunningham, C.W., 1996. Combining data in phylogenetic analysis. Trends Ecol. Evol. 11, 152–158.
Ingman, M., Kaessmann, H., Pääbo, S., Gyllensten, U., 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408, 708–713.
Janke, A., Arnason, U., 1997. The complete mitochondrial genome of Alligator mississippiensis and the separation between recent archosauria (birds and crocodiles). Mol. Biol. Evol. 14, 1266–1272.
Jermann, R.M., Opitz, J.G., Stackhouse, J., Benner, S.A., 1995. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. Nature 374, 57–59.
Jun, S.R., Sims, G.E., Wu, G.A., Kim, S.H., 2010. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: an alignment-free method with optimal feature resolution. Proc. Natl. Acad. Sci. U. S. A. 107, 133–138.
Kantorovitz, M.R., Robinson, G.E., Sinha, S., 2007. A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics 23, i249–i255.
Korf, I.F., Rose, A.B., 2009. Applying word-based algorithms: the IMEter. Methods Mol. Biol. 553, 287–301.
Kullberg, M., Nilsson, M., Arnason, U., Harley, E.H., Janke, A., 2006. Housekeeping genes for phylogenetic analysis of eutherian relationships. Mol. Biol. Evol. 23, 1493–1503.
Liu, F.G., Miyamoto, M.M., Freire, N.P., Ong, P.Q., Tennant, M.R., Yong, T.S., Gugel, K.F., 2001. Molecular and morphological supertrees for eutherian (placental) mammals. Science 291, 1786–1789.
Maddison, D.R., Ruvolo, M., Swofford, D.L., 1992. Geographic origins of human mitochondrial DNA: phylogenetic evidence from control region sequences. Syst. Biol. 41, 111–124.
Nei, M., 1996. Phylogenetic analysis in molecular evolutionary genetic. Annu. Rev. Genet. 30, 371–403.
Olivio, P.D., VandeWalle, M.J., Laipis, P.J., Hauswirth, W.W., 1983. Nucleotide sequence evidence for rapid genotypic shifts in the bovine mitochondrial DNA D-loop. Nature 306, 400–402.
Ota, T., Nei, M., 1994. Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. Mol. Biol. Evol. 11, 469–482.
Qi, J., Wang, B., Hao, B.L., 2004. Whole proteome prokaryote phylogeny without sequence alignment: a k-string comparison approach. J. Mol. Evol. 58, 1–11.
Raina, S.Z., Faith, J.J., Disotell, T.R., Seligmann, H., Stewart, C.B., Pollock, D.D., 2005. Evolution of base-substitution gradients in primate mitochondrial genomes. Genome Res. 15, 665–673.
Rzhetsky, A., Nei, M., 1992. A simple method for estimating and testing minimum-evolution tree. Mol. Biol. Evol. 9, 945–967.
Seutin, G., Lang, B.F., Mindell, D.P., Morais, R., 1994. Evolution of the WANCY region in amniote mitochondrial DNA. Mol. Biol. Evol. 11, 329–340.
Sims, G.E., Jun, S.R., Wu, G.A., Kim, S.H., 2009a. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc. Natl. Acad. Sci. U. S. A. 106, 2677–2682.
Sims, G.E., Jun, S.R., Wu, G.A., Kim, S.H., 2009b. Whole-genome phylogeny of mammals: evolutionary information in genic and non-genic regions. Proc. Natl. Acad. Sci. U. S. A. 106, 17077–17082.
Stuart, G.W., Berry, M.W., 2004. An SVD-based comparison of nine whole eukaryotic genomes supports a coelomate rather than ecdysozoan linkage. BMC Bioinforma. 5, 204.
Stuart, G.W., Moffett, K., Leader, J.J., 2002. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. Mol. Biol. Evol. 19, 554–562.
Takahata, N., 1993. Allelic genealogy and human evolution. Mol. Biol. Evol. 10, 2–22.
Tamura, K., Nei, M., 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10, 512–526.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.
Wen, J., Zhang, Y.Y., 2009. A 2D graphical representation of protein sequence and its numerical characterization. Chem. Phys. Lett. 476, 281–286.
Wistow, G., 1993. Lens crystallins: gene recruitment and evolutionary dynamism. Trends Biochem. Sci. 18, 301–306.
Wu, T.J., Burke, J.P., Davison, D.B., 1997. A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. Biometrics 53, 1431–1439.
Wu, T.J., Hsieh, Y.C., Li, L.A., 2001. Statistical measures of DNA dissimilarity under Markov chain models of base composition. Biometrics 57, 441–448.
Wu, T.J., Huang, Y.H., Li, L.A., 2005. Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences. Bioinformatics 21, 4125–4132.
Xia, X., Xie, Z., Kjer, K.M., 2003. 18S ribosomal RNA and tetrapod phylogeny. Syst. Biol. 52, 283–295.
Yang, X.W., Wang, T.M., 2013. A novel statistical measure for sequence comparison on the basis of k-word counts. J. Theor. Biol. 318, 91–100.
Yu, H.J., 2013. Segmented K-mer and its application on similarity analysis of mitochondrial genome sequences. Gene 518, 419–424.
Yu, Z.G., Zhou, L.Q., Anh, V., Chu, K.H., Long, S.C., Deng, J.Q., 2005. Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on all protein sequences from whole genome without sequence alignment. J. Mol. Evol. 60, 538–545.
Yu, C., Liang, Q., Yin, C., He, R.L., Yau, S.S., 2010. A novel construction of genome space with biological geometry. DNA Res. 17, 155–168.
Zardoya, R., Meyer, A., 1998. Complete mitochondrial genome suggests diapsid affinities of turtles. Proc. Natl. Acad. Sci. U. S. A. 95, 14226–14231.