BIOLOGICAL SCIENCES: Biophysics and Computational Biology

The *k*-mer natural vector and its application to the phylogenetic analysis of genetic sequences

Jia Wen[a,b], Raymond H. Chan[b], Rong L. He[c], and Stephen S.-T. Yau[d,1]

[a]School of Information Science, Suihua University, Suihua, PR China; [b]Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong; [c]Department of Biological Sciences, Chicago State University, Chicago, IL, USA; and [d]Department of Mathematical Science, Tsinghua University, Beijing, PR China

[1]To whom correspondence should be addressed.

Tel.: + 86 10 62787874; fax: + 86 10 62798033.

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

k-mer model | natural vector | phylogenetic analysis

Phylogenetic analysis of genetic sequences is very important for studying and exploring the evolutionary relationships of all types of individual organisms. However, most methods for phylogenetic analysis are alignment-based, assume some sort of evolutionary model, involve high computational complexity, and yield results that are often controversial. Therefore, based on the well-known *k*-mer model, we propose a *k*-mer natural vector model for representing a genetic sequence based on the number and distribution 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. This makes it more effective for handling whole or partial genomes than sequence alignment methods. Our proposed method is applied to the phylogenetic analysis of mitochondrial genome sequences and 18S rRNA sequences. The results of this phylogenetic analysis seem to be more reasonable than what has been produced by sequence alignment and by some published papers based on correlative specialized knowledge, and the computation time for this process is much lower than for sequence alignment. All results from applying this method to real datasets have demonstrated that the *k*-mer natural vector method is a very powerful tool for analyzing and annotating genetic sequences and determining phylogenetic relationships both in terms of accuracy and efficiency.

Significance statement:

The *k*-mer natural vector method is a very powerful tool for performing phylogenetic analysis of genetic sequences (DNA and RNA). For a given value of *k*, the *k*-mer natural vector is able to represent the information of *k*-mers in the sequence. Our proposed method for the phylogenetic analysis of genetic sequences does not require an evolutionary model or human intervention. Whole or partial genome datasets can be more easily and more quickly analyzed and annotated with our proposed method than by using traditional alignment-based methods. Our method is not only more accurate in determining evolutionary relationships between genetic sequences, but it also greatly reduces the computation time required for phylogenetic analysis.

\body

Phylogenetic analysis of genetic sequences has become essential for researching the evolutionary relationships between all types of individual organisms (from bacteria to humans) (1). Phylogenetic analysis is also important for clarifying the evolutionary pattern of multigene families (2-4), as well as for understanding adaptive evolution at the molecular level (5-7). It also provides deep insight into the mechanism for the maintenance of polymorphic alleles in populations (8, 9). The results of phylogenetic analysis are represented by phylogenetic trees, 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, compensate for the ineffectiveness of traditional alignment-based methods.

Among all alignment-free methods, the *k*-mer 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 (10), and the counts of *k*-mers appearing in the sequence were used for the comparison of regulatory sequences by Kantorovitz et al. (11). Later, various frequency-based methods were introduced for sequence comparison presented by Wu et al. (12-14), Korf and Rose (15), Sims et al. (16, 17), and Jun et al. (18).

The advantage of the *k*-mer 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 (19, 20).

The natural vector approach (21) is an alternate alignment-free method which produces a one-to-one association between genetic sequences and vectors in a finite dimensional space. 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 the genetic sequences.

In this paper, we demonstrate how these two approaches can be combined to produce a *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 the *k*-mers appearing in the sequence.

Results

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

As the first application, we analyzed the mitochondrial genome sequences of 31 species using our method. This data was previously analyzed using the original natural vector method (21). The descriptions of the 31 mitochondrial genome sequences are listed in the Table S1 of Supporting Information (SI), the lengths of which are from 16338 to 17447 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 (22, 23). The phylogenetic tree of 31 mitochondrial genomes obtained by our proposed method is shown in Figure 1.

Looking at Figure 1, all 31 genomes are correctly clustered into eight known clusters: Carnivora (red), Perissodactyla (blue), Artiodactyla (yellow), Cetacea (brown), Primates (green), Lagomorpha (light blue), Rodentia (purple), and Erinaceomorpha (light green). Since whales evolved from the primitive artiodactyl, the 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 cluster with dormouse and squirrel, in that, they are all glires. The resulting phylogenetic tree agrees with those from standard biological taxonomy, sequence alignment, and some published papers (20-26). Compared with Figure 3 of (21) drawn by original natural vector method, the accuracy of evolutionary relationships has been improved, which can be easily seen from the evolutionary relationships within the subgroups of Primates and Carnivora, respectively.

We also applied our method to investigate variations in human mitochondrial genomes and to explore the origin of modern humans. Mitochondrial DNA (mtDNA) has been a potential tool in

the study of human evolution, owing to characteristics such as high copy number, lack of recombination (27), high substitution rate (28), and maternal mode of inheritance (29). Most studies of human evolution based on the mtDNA have been confined to the control region, which constitutes less than 7% of the mitochondria genome. These studies are complicated by the extreme variation in substitution rates between sites, and parallel mutations causing difficulties in the estimation of genetic distance. These factors have made phylogenetic inferences questionable (30, 31).

To improve the information obtained from mitochondrial DNA for studies of human evolution, Ingman et al. described the global mtDNA diversity in humans based on the sequence alignment of complete mtDNA sequences (excluding D-loops) from 53 diverse origins (32). 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 15440 to 15450 base pairs (bp). They are described in the Table S2 of SI, and the phylogenetic tree for them that was obtained using our proposed method is shown in Figure 2.

From Figure 2, 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 accordance with known evidence of human evolution and human migration. Compared with Figure 2 of (32), the evolutionary relationships between all Africans and most non-Africans are the same, and differences only exist in several non-Africans.

Moreover, the results of our proposed method seem to be more reasonable than what has been produced by sequence alignment method. For example, sequence alignment method would imply that two mtDNA samples from Australians (Australians1 and Australians2) were closely connected, but our method (see Figure 2) shows the contrary. If we take English and Crimean Tatar as references, the lengths of four mtDNAs considered are all 15449. The mismatches between Australian1 and Australian2, English, Crimean Tatar are 20, 12, and 16, respectively, and mismatches between Australian2 and Australian1, English, Crimean Tatar all equal 20. Hence, the two Australians should not close connected in phylogenetic tree, and phylogenetic tree obtained by our method looks more reasonable.

Additionally, our method was used to analyze the phylogeny of tetrapod 18S rRNA sequences. 18S rRNA has consequently been considered odd, providing significantly different estimates of phylogeny in higher organisms (33). The phylogenetic relationship amongst 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. Previous phylogenetic analyses of tetrapod 18S rRNA sequences seemed to support the grouping of birds and mammals (34), whereas other molecular data and morphological and paleontological data favor the grouping of birds with crocodilians (35) which is more acceptable to biologists. We have investigated this question by applying our method to the tetrapod dataset shown in Figure 3 of (36) which contains sequences whose lengths are from 1733 to 2235 base pairs (bp).

The phylogenetic tree based on our method was shown in Figure 3. This 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 methods and what is found in some phylogenetic analyses (34-44). It

can be seen that birds group with crocodilians rather than group with mammals. This result conforms to traditional classification and the results in (35, 36). Compared with Figure 3 of (36), our phylogenetic tree is better, in that, homo and oryctolagus are closer in our figure, which fits the results of sequence alignment.

To further show the utility of our *k*-mer natural vector method, we perform multiple sequence alignment on the same datasets that we considered above, using the MEGA 5 implementation of the clusterW algorithm. The phylogenetic trees based on the sequence alignment are shown in Figure 4-6, where the species are colored the same as Figure 1-3, respectively. Here, we only consider the differences between corresponding phylogenetic trees constructed by the *k*-mer natural vector method and clusterW, respectively. The configuration of our computer is Intel(R) Core(TM) i5-2450M CPU @2.50GHz 2.50 GHz Microsoft Windows 7 with 4.00G RAM.

When clusterW was applied to the 31 mitochondrial genome sequences, rhinoceroses seem close the carnivore in Figure 4, rather than the cetartiodactyla in Figure 1, which does not agree with standard biological taxonomy, in that, rhinoceros is graminivorous. The computation time for our *k*-mer natural vector method is 190.613091 seconds, which is less than the time of multiple sequence alignment by clusterW (about 4.75 hours).

We also applied multiple sequence alignment to the 53 human mtDNA sequences. The phylogenetic tree obtained is shown in Figure 5. This tree is the same as one shown in Figure 2 of (32) obtained by sequence alignment. As discussed above, our *k*-mer natural vector method seems to get a better result. Moreover, the computation time of our proposed method is 68.244747 seconds that is far less than the time of clusterW (more than 11 hours).

Finally, we applied clusterW to the tetrapod 18S rRNA sequences. Our phylogenetic tree is better, because birds group with crocodilians in Figure 3, rather than mammals shown in Figure 6, which conforms to traditional classification (35, 36). The time of our method is 6.716700 seconds, and the time of multiple sequence alignment is about 5 minutes.

Discussion

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 components of the *k*-mer natural vector, which contains information of relationships between *k*-mers in a sequence. The correspondence between a genetic sequence and its 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 the high computational complexity associated with sequence alignment. Its application to real datasets has shown that the *k*-mer natural vector method is a powerful tool for the phylogenetic analysis of genetic sequences. It not only enhances the accuracy of evolutionary relationships, but it also greatly reduces the computation time for phylogenetic analysis.

Although our improved natural vector method has proved itself useful in the phylogenetic analysis for most genetic sequences, it may not be as effective for some special cases where the sequences lengths are different from what we have considered.

Materials and Methods

***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, $'N_1N_2\cdots N_L'$, where $N_l \in \{A, C, G, T\}$, $l = 1,2,\cdots, L$. A string of consecutive *k* nucleotides within the genetic sequence is called a *k*-mer. The *k*-mers appearing in a sequence may be enumerated be 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 $n^{(s,k)} = (n_{s[1]}, n_{s[2]}, \cdots, n_{s[4^k]})$ where $n_{s[i]}$ is the number of times the *k*-mer $[i]$ occurs in the sequence.

***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 :

(1) The *k*-mer counting vector $n^{(s,k)}$ as defined above.

(2) The *k*-mer mean distance vector $(\mu_{[1]}, \mu_{[2]}, \dots, \mu_{[4^k]})$, where $\mu_{[i]}$ is defined to be the arithmetic mean of the distances from the various 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.

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

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

where $n_{[i]}$ denotes the number of times $[i]$ appears 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 the genetic sequence, and $\mu_{[i]}$ is the mean of the 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 to *k*-mer distribution within the sequence.

When *k*=1, the *k*-mer natural vector is the same to the original natural vector. Thus the *k*-mer natural vector 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 may not be 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 natural vector is one-to-one for each given *k*, which is similar to the proofs in (21, 45). Because all the first central moments are zero, we do not need to include them as part of the natural vector.

Because the 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, the natural vector contains information about the relationships of *k*-mers. Because of this, the *k*-mer natural vector model overcomes the deficiency of *k*-mer model.

It has been verified that it is not necessary to include normalized central moments higher than

second order for the comparison of DNA and protein sequences (45, 46), so using a 3×4^k-dimensional vector $(n_{[i]}, \mu_{[i]}, D_2^{[i]})$ is enough to represent a genetic sequence.

The choice of *k*. Because the parameter *k* has a great influence on the results of phylogenetic analysis and on the complexity of computation, it is very important to choose a suitable *k*. Some researchers have explored the selection of *k*. 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 (14). Another investigation was done by Sims et al. (16), who reported that the optimal length of word lies within an approximate range with lower bound $\log_4 n$ and the upper bound given by the criterion that phylogenetic tree topology for length *k* must be parallel to that of $k + 1$. Based on their work, we chose the value of *k* to be within a range $[\mathrm{floor}(\log_4 \min(L)), \mathrm{ceil}(\log_4 \max(L))]$, where L is the set of lengths of genetic sequences being considered. This explicit range greatly shortens the range of values of *k* that need to be considered.

Distance metric. Since each genetic sequence can be uniquely represented by a 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 eliminates the effects of high dimensionality (47, 48). In this paper, we select the genetic distance defined below to measure the similarities of genetic sequences, which has been widely used in the *k*-mer model (49-51).

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

$$d(\mathrm{s}_1, \mathrm{s}_2) = 1 - \cos(v_1, v_2) = 1 - \frac{v_1 \cdot v_2}{|v_1||v_2|},$$

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 vector v_1, v_2, respectively.

Once genetic distances among all genetic sequences are obtained, the evolutionary tree can be determined by the neighbor-joining (NJ) method using MEGA 5 (52, 53).

Supporting Information (SI)

SI contains the proof that the correspondence between a genetic sequence and its natural vector is one-to-one for each given *k* and Genbank ID information for 31 mitochondrial genomes and 53 human mtDNAs.

ACKNOWLEDGEMENTS. We thank Dr. Max Benson for critically reading and editing the manuscript. This work is supported by Natural and Technology Research Project of Heilongjiang Provincial Education Department (12513097), Youth Foundation of Suihua University (KQ1202004), U.S. NSF grant DMS-1120824, China NSF grant 31271408, and Tsinghua University.

Footnotes

Author contributions: S.Y. and J.W. designed research; S.Y. and J.W. performed research; and J.W., S.Y., R.C., and R.H. wrote the paper.

The authors declare no conflict of interest.

References

1. Nei M (1996) Phylogenetic analysis in molecular evolutionary genetic. *Annu Rev Genet* 30:371–403.
2. Atchley WR, Fitch WM, Bronner FM (1994) Molecular evolution of the MyoD family of transcription factors. *Proc Natl Acad Sci USA* 91(24):11522–11526.
3. Goodwin RL, Baumann H, Berger FG (1996) Patterns of divergence during evolution of α_1-proteinase inhibitors in mammals. *Mol Biol Evol* 13(2):346–358.
4. 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(3):469–482.
5. Chandrasekharan UM, Sanker S, Glynias MJ, Karnik SS, Husain A (1996) Angiotensin II-forming activity in a reconstructed ancestral chymase. *Science* 271(5248):502–505.
6. Jermann RM, Opitz JG, Stackhouse J, Benner SA (1995) Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* 374(6517):57–59.
7. Wistow G (1993) Lens crystallins: gene recruitment and evolutionary dynamism. *Trends Biochem Sci* 18(8):301–306.
8. Figueroa F, Gunther E, Klein J (1988) MHC polymorphism pre-dating speciation. *Nature* 335(6187):265–267.
9. Takahata N (1993) Allelic genealogy and human evolution. *Mol Biol Evol* 10(1):2–22.
10. Blaisdell BE (1986) A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc Natl Acad Sci USA* 83(14):5155–5159.
11. Kantorovitz MR, Robinson GE, Sinha S (2007) A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics* 23(13):i249–i255.
12. Wu TJ, Burke JP, Davison DB (1997) A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. *Biometrics* 53(4):1431–1439.
13. Wu TJ, Hsieh YC, Li LA (2001) Statistical measures of DNA dissimilarity under Markov chain models of base composition. *Biometrics* 57(2):441–448.
14. Wu TJ, Huang YH, Li LA (2005) Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences. *Bioinformatics* 21(22):4125–4132.
15. Korf IF, Rose AB (2009) Applying word-based algorithms: the IMEter. *Methods Mol Biol* 553:287–301.
16. Sims GE, Jun SR, Wu GA, Kim SH (2009) Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc Natl Acad Sci USA* 106(8):2677–2682.
17. Sims GE, Jun SR, Wu GA, Kim SH (2009) Whole-genome phylogeny of mammals: evolutionary information in genic and non-genic regions. *Proc Natl Acad Sci USA* 106(40):17077–17082.
18. Jun SR, Sims GE, Wu GA, Kim SH, (2010) Whole-proteome phylogeny of prokaryotes by feature frequency profiles: an alignment-free method with optimal feature resolution. *Proc Natl Acad Sci USA* 107(1):133–138.
19. Yang XW, Wang TM (2013) A novel statistical measure for sequence comparison on the basis of *k*-word counts. *J Theor Biol* 318:91-100.
20. Yu HJ (2013) Segmented *K*-mer and its application on similarity analysis of mitochondrial genome sequences. *Gene* 518(2):419-424.
21. Deng M, Yu C, Liang Q, He RL, and Yau SS (2011) A novel method of characterizing genetic sequences: genome space with biological distance and applications. *PLoS One* 6(3): e17293.
22. Yu C, Liang Q, Yin C, He RL, Yau SS (2010) A Novel Construction of Genome Space with Biological Geometry. *DNA Res* 17(3):155-168.
23. Huang G, Zhou H, Li YF, Xu L (2011) Alignment-free comparison of genome sequences by a new numerical

characterization. *J Theor Biol* 281(1):107-112.

24. Liu FG, Miyamoto MM, Freire NP, Ong PQ, Tennant MR, Yong TS, Gugel KF (2001) Molecular and morphological supertrees for eutherian (placental) mammals. *Science* 291(5509):1786-1789.
25. Raina SZ, Faith JJ, Disotell TR, Seligmann H, Stewart CB, Pollock DD (2005) Evolution of base-substitution gradients in primate mitochondrial genomes. *Genome Res* 15(5):665-673.
26. Kullerg M, Nilsson M, Arnason U, Harley EH, Janke A (2006) Housekeeping genes for phylogenetic analysis of eutherian relationships. *Mol Biol Evol* 23(8):1493-1503.
27. Olivio PD, VandeWalle MJ, Laipis PJ, Hauswirth WW (1983) Nucleotide sequence evidence for rapid genotypic shifts in the bovine mitochondrial DNA D-loop. *Nature* 306(5941):400-402.
28. Brown WM, George MJ, and Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. *Proc Natl Acad Sci USA* 76(4):1967-1971.
29. Giles RE, Blanc H, Cann HM, and Wallace DC (1980) Maternal inheritance of human mitochondrial DNA. *Proc Natl Acad Sci USA* 77(11):6715-6719.
30. 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(3):512-526.
31. Maddison DR, Ruvolo M, and Swofford DL (1992) Geographic origins of human mitochondrial DNA: phylogenetic evidence from control region sequences. *Syst Biol* 41:111-124.
32. Ingman M, Kaessmann H, Pääbo S, and Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. *Nature* 408(6813):708-713.
33. Huelsenbeck JP, Bull JJ, and Cunningham CW (1996) Combining data in phylogenetic analysis. *Trends Ecol Evol* 11(4):152-158.
34. Xia X, Xie Z, and Kjer KM (2003) 18S ribosomal RNA and tetrapod phylogeny. *Syst Biol* 52(3):283-295.
35. Hedges SB, Moberg KD, and Maxson LR (1990) Tetrapod phylogeny inferred from 18S and 28S ribosomal RNA sequence and a review of the evidence for amniote relationships. *Mol Biol Evol* 7(6):607-633.
36. Chan RH, Chan TH, Yeung HM, and Wang RW (2011) Composition vector method based on maximum entropy principle for sequence comparison. *IEEE/ACM Trans Comput Biol Bioinform* 9:79-87.
37. Rzhetsky A, Nei M (1992) A simple method for estimating and testing minimum-evolution tree. *Mol Biol Evol* 9(5):945-967.
38. Hedges SB (1994) Molecular evidence for the origin of birds. *Proc Natl Acad Sci USA* 91(7):2621-2624.
39. Seutin G, Lang BF, Mindell DP, and Morais R (1994) Evolution of the WANCY region in amniote mitochondrial DNA. *Mol Biol Evol* 11(3):329-340.
40. Caspers GJ, Reinders GJ, Leunissen JA, Wattel J, and Dejong WW (1996) Protein sequences indicate that turtles branched off from the amniote tree after mammals. *J Mol Evol* 42(5):580-586.
41. 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(12):1266-1272.
42. Zardoya R, Meyer A (1998) Complete mitochondrial genome suggests diapsid affinities of turtles. *Proc Natl Acad Sci USA* 95(24):14226-14231.
43. Ausio J, Soley JT, Burger W, Lewis JD, Barreda D, and Cheng KM (1999) The histidine-rich protamine from ostrich and tinamou sperm: A link between reptile and bird protamines. *Biochemistry* 38(1):180-184.
44. Dixon, MT, Hillis DM (1993) Ribosomal RNA secondary structure: Compensatory mutations and implications for phylogenetic analysis. *Mol Biol Evol* 10(1):256-267.
45. Yu C, Deng M, Cheng SY, Yau SC, He RL, and Yau SS (2013) Protein space: a natural method for realizing the nature of protein universe. *J Theo Biol* 318:197-204.
46. Yu C, Hernandez T, Zheng H, Yau SC, Huang HH, He RL, Yang J, and Yau, SS (2013) Real time classification of

viruses in 12 dimensions. *PLoS ONe* 8(5): e64328.

47. Berry MW, Drmac Z, and Jessup ER (1999) Matrices, vector spaces, and information retrieval. *SIAM Rewiew* 41(2):335-362.
48. Wen J, Zhang YY, (2009) A 2D graphical representation of protein sequence and its numerical characterization. *Chem Phys Lett* 476:281-286.
49. Qi, J, Wang B, and Hao BL (2004) Whole proteome prokaryote phylogeny without sequence alignment: a k-string comparison approach. *J Mol Evol* 58(1):1-11.
50. Stuart GW, Moffett K, and Leader JJ (2002) A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Mol Biol Evol* 19(4):554-562.
51. Stuart GW, Berry MW (2004) An SVD-based comparison of nine whole eukaryotic genomes supports a coelomate rather than ecdysozoan linkage. *BMC Bioinformatics* 5:204.
52. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4(4):406-425.
53. Studier JA, Keppler KJ (1988) A note of the neighbor-joining algorithm of Saitou and Nei. *Mol Biol Evol* 5(6):729-731.

Figure legends

Figure 1. Phylogenetic tree of 31 mitochondrial genome sequences based on 7-mer natural vector. All 31 genomes are correctly clustered into eight known clusters: Carnivora (red), Perissodactyla (blue), Artiodactyla (yellow), Cetacea (brown), Primates (green), Lagomorpha (light blue), Rodentia (purple), and Erinaceomorpha (light green), which are consistent with results of taxonomy in biology.

Figure 2. Phylogenetic tree of 53 human mitochondrial genome sequences based on 6-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 are accordance with evidences of human evolution and human migration.

Figure 3. Phylogenetic tree of 40 18S rRNA sequences based on 5-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 the data from morphology and paleontology, and traditional classification of biology.

Figure 4. Phylogenetic tree of 31 mitochondrial genome sequences based on sequence alignment (clusterW).

Figure 5. Phylogenetic tree of 53 human mitochondrial genome sequences based on sequence alignment (clusterW).

Figure 6. Phylogenetic tree of 40 18S rRNA sequences based on sequence alignment (clusterW).