Genomics 111 (2019) 1298–1305

Contents lists available at ScienceDirect

Genomics

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

Methods Paper

Phylogenetic analysis of protein sequences based on a novel *k*-mer natural vector method

YuYan Zhang[a,1], Jia Wen[b,*,1], Stephen S.-T. Yau[c,*]

[a] School of Agriculture and hydraulic Engineering, Suihua University, Suihua 152061, China
[b] School of Information Engineering, Suihua University, Suihua 152061, China
[c] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Keywords:
Phylogenetic analysis
Protein sequence
k-mer model
Natural vector
Neighbor joining

ABSTRACT

Based on the *k*-mer model for protein sequence, a novel *k*-mer natural vector method is proposed to characterize the features of *k*-mers in a protein sequence, in which the numbers and distributions of *k*-mers are considered. It is proved that the relationship between a protein sequence and its *k*-mer natural vector is one-to-one. Phylogenetic analysis of protein sequences therefore can be easily performed without requiring evolutionary models or human intervention. In addition, there exists no a criterion to choose a suitable *k*, and *k* has a great influence on obtaining results as well as computational complexity. In this paper, a compound *k*-mer natural vector is utilized to quantify each protein sequence. The results gotten from phylogenetic analysis on three protein datasets demonstrate that our new method can precisely describe the evolutionary relationships of proteins, and greatly heighten the computing efficiency.

1. Introduction

Phylogenetic analysis is the study of evolutionary relationships among molecules, phenotypes, and organisms [1]. In the context of protein sequence data, phylogenetic analysis is the key cornerstone of comparative sequence analysis and has many applications in the study of protein evolution and functions, as well as genome annotation, gene function prediction, identification and construction of gene families, and gene discovery [2]. Therefore, using protein sequences to analyze the phylogeny of species makes more sense than using DNA sequences [3–5]. Proteins with high sequence identity tend to possess similarity in function and evolutionary relationship, and results obtained from phylogenetic analysis are represented by a phylogenetic tree, in which sequences are grouped based on sequence similarities.

With the rapid increase of sequence data in the past decades, plenty of approaches have been proposed for protein phylogenetic analysis [6–12]. Most of them depend on multiple sequence alignment, which commonly assumes some sort of evolutionary model, yielding disagreement of interpretation. Although alignment-based methods achieve satisfactory results in evolutionary relationships, they often involve in high computational complexity. Notably, some of them break down when fed whole genome data. Additionally, as in the case of viral genomes, several fail to handle gene rearrangements issues. Hence, alignment-free methods based on the numerical characterizations of biological sequences were proposed to improve alignment-based methods.

Many biological molecular studies using *k*-mer model methods have already appeared [13–16]. The most significant advantage for *k*-mer model methods is that a phylogenetic tree can be constructed much faster. However, sequence relationships are more or less neglected. The original natural vector approach was proposed to incorporate the normalized central moments to account for the interrelationships between different portions of genetic sequences, and produced one-to-one relationship between genetic sequences and vectors in a finite dimensional space [17,18]. But the obtaining results cannot accurately describe the phylogeny of species [19]. Recently, He et al. [5] proposed a feature vector to describe the composition of amino acids in a protein sequence.

In this paper, we propose a simple but efficient *k*-mer natural vector method to numerically characterize a protein sequence, utilizing the frequencies and positional information of *k*-mers in a protein sequence. The obtaining results have shown that our new *k*-mer natural vector method can offer a credible phylogeny depicting the evolutionary relationship of species.

* Corresponding authors.
E-mail addresses: wenjia198021@126.com (J. Wen), yau@uic.edu (S.S.-T. Yau).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ygeno.2018.08.010
Received 25 July 2018; Received in revised form 19 August 2018; Accepted 27 August 2018
Available online 05 September 2018

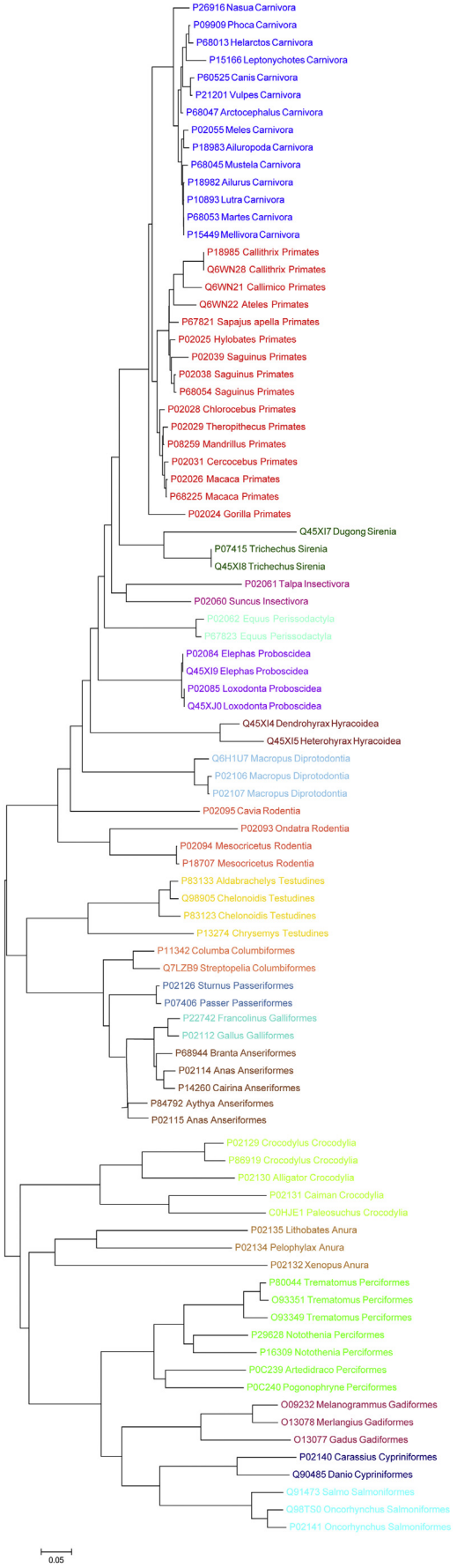

Fig. 1. NJ phylogenetic tree of beta-globin protein sequence of 88 species based on the *k*-mer natural vector method. The 88 beta-globin sequences are correctly clustered into 20 groups: Carnivora, Primates, Sirenia, Insectivora, Perissodactyla, Hyracoidea, Proboscidea, Rodentia, Diprotodontia, Testudines, Columbiformes, Passeriformes, Galliformes, Anseriformes, Crocodylia, Anura, Perciformes, Gadiformes, Cypriniformes, and Salmoniformes. This resulting phylogenetic tree agrees well with the results in standard biological taxonomy and the evolutionary relationship of species.

2. Materials and methods

2.1. Dataset

Three sets of real protein sequence data are assembled and utilized to explore the evolutionary relationship of proteins, in which both long and short sequences are considered. Dataset S1 consists of 88 beta-globin sequences from different species, and dataset S2 is composed of 116 human rhinoviruses belonging to the Enterovirus genus in the Picornaviridae family. In addition, a much larger dataset S3 containing 1163 influenza A viruses isolated in China is also included.

2.2. K-mer model of protein sequence

The *k*-mer model for protein sequence mimics that for genetic sequence. There are 20 amino acids. Each has a name, a 3-letter shorthand name, or a single letter symbol. Thus, α = A indicates that the amino acid α is Alanine, whose 3-letter shorthand is Ala and α = W indicates Tryptohan, whose 3-letter shorthand is Trp. A protein sequence consists of amino acids linearly arranged.

Let $\boldsymbol{\varphi} = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ be the set of 20 amino aicds and let $s = < \alpha_1, \alpha_2, \cdots, \alpha_L >$ be a protein sequence sof length L, Lwhere $N_l \in \boldsymbol{\varphi}$, $l = 1, 2, \cdots, L$. $AN_l \in \{A, C, G, T\}$ *k*-mer is a string of *k* consecutive single letter symbols and numbered left to right. Given any positive integer *k*, there are 20^k different possible sequences, or rather 20^k different possible *k*-mers. Thus $L - k + 1$ *k*-mer are determined by sliding a window of width *k* along the entire sequence of length L. For $k \geq 2$, adjacent *k*-mers from such a sequence are highly correlated, less so the farther apart they are.

2.3. K-mer natural vector for protein sequence

For any protein sequence s and a given *k*, the *k*-mer natural vector is defined to be the concatenation of following three vectors, each of which is of length 20^k:

(1) The *k*-mer counting vector $(n_{s[1]}, n_{s[2]}, \cdots, n_{s[20^k]})$, where $n_{s[i]}$ is the number of *k*-mer $s[i]$ occurring in sequence s.
(2) The *k*-mer mean distance vector $(\mu_{s[1]}, \mu_{s[2]}, \cdots, \mu_{s[20^k]})$, where $\mu_{s[i]}$ is the arithmetic mean of the distances of the *k*-mer s[*i*] to the first base. If a specific *k*-mer $s[i]$ does not exist, $\mu_{s[i]}$ is defined to be zero.
(3) The *k*-mer normalized central moment vector $(D_2^{s[1]}, D_2^{s[2]}, \cdots, D_2^{s[20^k]})$, the component of which $(D_2^{s[i]})$ is the variance for the distances of *k*-mer $s[i]$ to the first base, which is defined as follows:

$$D_m^{[i]} = \sum_{j=1}^{n_{[i]}} \frac{(s[i][j] - \mu_{[i]})^m}{n_{[i]}^{m-1}(L - k + 1)^{m-1}}, \; m = 1, 2, \cdots, n_{[i]} D_2^{s[i]}$$

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

where $n_{s[i]} n_{[i]}$ denotes the number of *k*-mer $s[i]$ appearing in the sequence s of length *L*, $s[i][j]$ is the distance of *j*th *k*-mer $s[i]$ from the first base in s. This differs from the usual variance by the extra factor of $L - k + 1$ in the denominator and is closely related to the moment vector whose components are $\Sigma_{j=1}^{n_{s[i]}}(s[i][j] - \mu_{s[i]})^2$ with no denominator factors. More generally, for $m > 2$ there are also

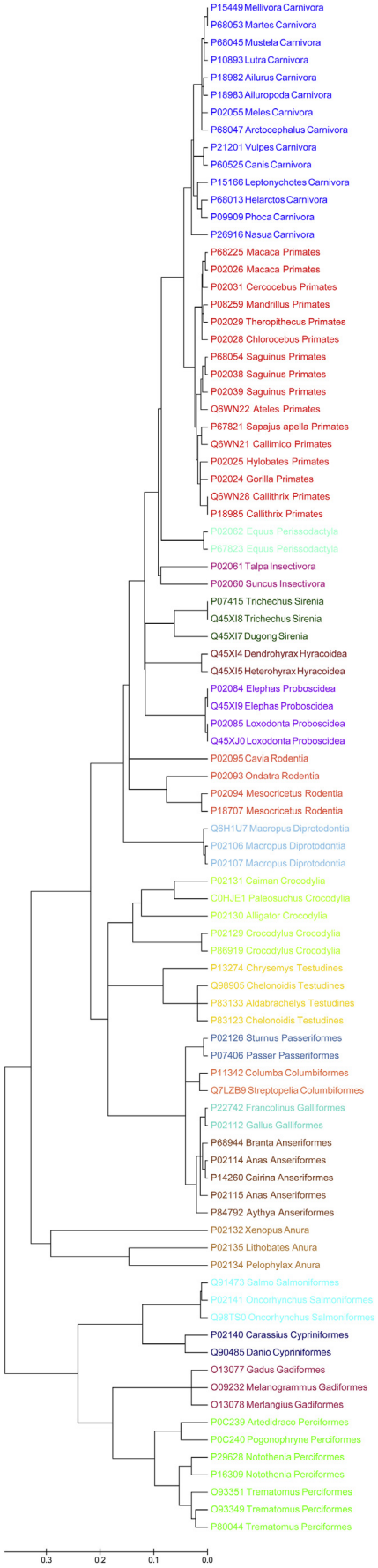

Fig. 2. NJ phylogenetic tree of beta-globin protein sequence of 88 species based on ClustalW. The 88 beta-globin sequences are clustered into 20 groups: Carnivora, Primates, Perissodactyla, Insectivora, Sirenia, Hyracoidea, Proboscidea, Rodentia, Diprotodontia, Crocodylia, Testudines, Passeriformes, Columbiformes, Galliformes, Anseriformes, Anura, Salmoniformes, Cypriniformes, Gadiformes, and Perciformes.

m^{th} moment vectors consisting of $\Sigma_{j=1}{}^{n_{s[i]}}(s[i][j]-\mu_{s[i]})^m$ and normalized moment vectors consisting of L $D_m^{s[i]}=\sum_{j=1}^{n_{s[i]}}\frac{(s[i][j]-\mu_{s[i]})^m}{n_{s[i]}{}^{m-1}\cdot(L-k+1)^{m-1}}$.

If the distributions of each k-mer are different, two protein sequences cannot be similar even though they contain the same set of k-mers and the same measurements for the total distance. The numerical parameters in each subset maybe not sufficient to annotate a protein sequence, but the combination of numerical parameters is sufficient to characterize each protein sequence. We mathematically prove that the relationship between a protein sequence and its corresponding k-mer natural vector is one-to-one for each given k in the Test S1.

K-mer natural vector is derived by concatenating the frequencies of occurrences of each k-mer in the sequence and its mean distance to the first base to the normalized central moments. Hence, k-mer natural vector contains information on relationships of k-mer, which is commonly neglected by former k-mer model methods.

It has shown that the $3*20^k3\times4^k$-dimensional vector $(n_{s[i]},\mu_{s[i]},D_2{}^{s[i]})$ $n_{[i]}$, $\mu_{[i]}$, $D_2{}^{[i]}$is enough to represent a protein sequence, and there is no necessary to include normalized central moments higher than second order, because the higher central moments hardly make any contribution. Hence, the $3*20^k$-dimensional natural vector mapping restricted on all the datasets is still one-to-one mapping.

2.4. A compound k-mer natural vector quantifying each protein sequence

The k-mer natural vector is proposed to describe the numbers and distributions of k-mers in a protein sequence, and the dimension of which is $3*20^k$ for each given k. In previous k-mer methods, there exists no a criterion to tell us how to choose k used. Obviously the parameter k has a great influence on obtaining results of evolutionary relationship. Specially, when k enlarges, the computation load would increase tremendously.

In this paper, we propose a compound k-mer natural vector (with $k=2$ and 3) is jointly to uniquely quantify each protein sequence. Therefore, each protein sequence can be numerically represented by a $2520=3\cdot(20^2+20^3)$ dimensional feature vector. In the section of Results and Discussion, it should be verified that both long and short protein sequences can be accurately depicted with the compound k-mer natural vector.

Once every protein sequence considered in the phylogenetic analysis is uniquely numerically characterized by a compound k-mer natural vector, the cosine distance metric can be utilized to calculate the pairwise distance of protein sequences, which has been widely used in k-mer model methods [20–23]. Then, the phylogenetic tree can be drawn through the method of Neighbor Joining (NJ) using MEGA 6.06 [24].

3. Results and discussion

To fully demonstrate the validity of the new method in depicting the evolutionary relationship of protein sequences, the k-mer natural vector method is applied in the phylogenetic analysis on beta-globin sequence, human rhinovirus, and influenza A virus, in which protein sequences of different lengths are considered.

3.1. Phylogenetic analysis of 88 beta-globin sequences

We first analyze 88 beta-globin sequences from different species, which is the most common haemoglobin in adult human and often utilized to explore the evolutionary relationships of species [17,25]. This dataset has been investigated by a new cluster method [5], and the variance in length is from 140 to 148. As a comparison, the NJ tree of 88 beta-globin sequences is shown in Fig. 1, using our novel *k*-mer natural vector method.

Look at Fig. 1, 88 beta-globin sequences are correctly clustered into 20 groups: Carnivora, Primates, Sirenia, Insectivora, Perissodactyla, Hyracoidea, Proboscidea, Rodentia, Diprotodontia, Testudines, Columbiformes, Passeriformes, Galliformes, Anseriformes, Crocodylia, Anura, Perciformes, Gadiformes, Cypriniformes, and Salmoniformes, which are the same to the results of [5]. Perciformes, Gadiformes, Cypriniformes, and Salmoniformes are all Teleosts, they group together, which conforms to the conclusion in Cladistic analysis [26]. In addition, Columbiformes, Passeriformes, Galliformes, and Anseriformes are belonging to the Galloanserae, the main group of modern birds [27]. Their clusters are supported with the morphological data and DNA sequence data [28,29]. Our resulting phylogenetic tree agrees well with those in standard biological taxonomy, and evolutionary relationship of species.

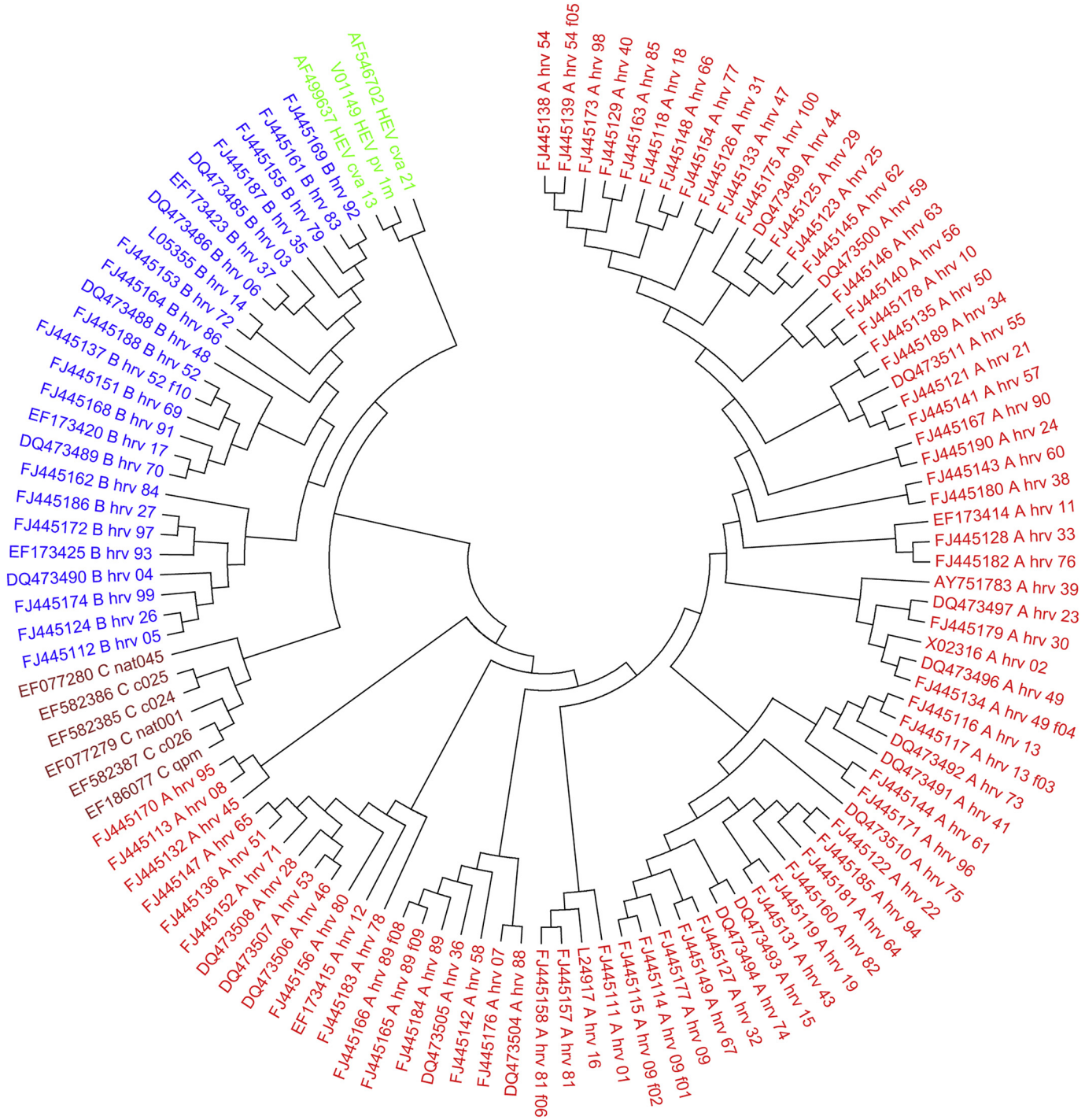

Fig. 3. NJ phylogenetic tree of 114 HRV serotypes based on the *k*-mer natural vector method. All 113 HRVs are clustered into three groups: HRV-A, HRV-B, and HRV-C, and 3 HEV-Cs form an outgroup, which are in according with clinical heterogeneity of HRV infections in humans and results gotten from published methods.

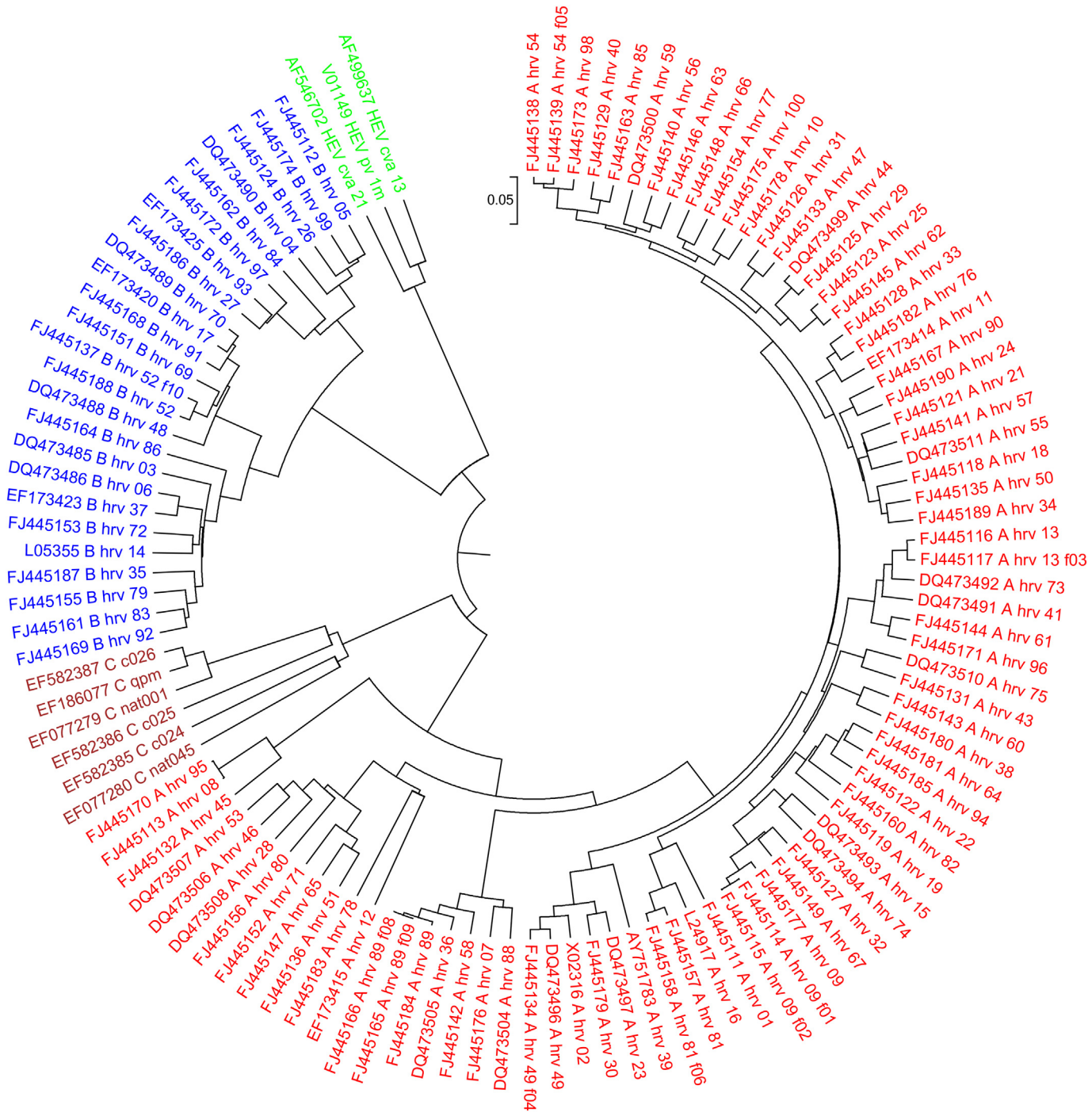

Fig. 4. NJ phylogenetic tree of 114 HRV serotypes based on ClustalW. All 113 HRVs are clustered into three groups: HRV-A, HRV-B, and HRV-C, and 3 HEV-Cs form an outgroup.

However, these results cannot be gotten by a new cluster method of [5].

To further show the utility of our new method, we also perform multiple sequence alignment on the same dataset, using MEGA 6.06 to execute algorithms of ClustalW and MUSCLE that are current most classic methods in the phylogenetic analysis. The phylogenetic trees of 88 beta-globin sequences drawn from ClustalW and MUSCLE are shown in Fig. 2 and Fig. 2*s* by NJ method, respectively, where species are coloured the same as in Fig. 1. Comparing Fig. 1 with Fig. 2, the evolutionary relationships of 88 beta-globin sequences are consistent with each other.

3.2. Phylogenetic analysis of 116 human rhinoviruses

Human rhinovirus (HRV), first discovered in the 1950s, is one of the most important causes of respiratory infections and has been associated mostly with the common cold [30]. The HRVs comprise the species of RV-A, RV-B, and RV-C Enterovirus genus in the Picornaviridae family, but the classification status is not always the case [31]. Meanwhile, the phylogenetic analysis of whole genome HRV genomic sequences show that the HRVs can be classified into three distinct groups, HRV-A, HRV-

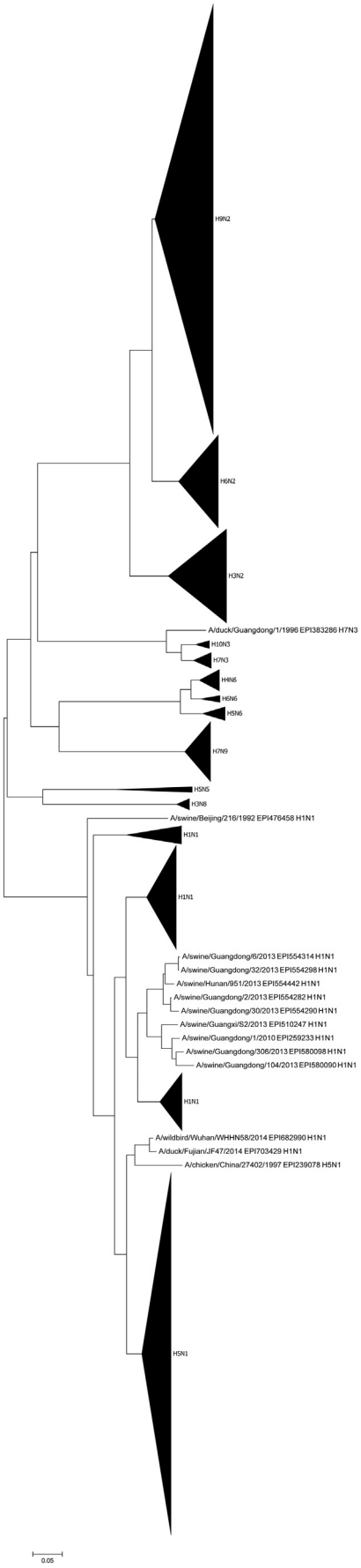

Fig. 5. NJ phylogenetic tree of 1163 influenza A viruses based on the *k*-mer natural vector method. All 1163 influenza A viruses are divided into two subgroups: one consisting the types of N2, N3, N6, and N9; and the other one containing the types of N8, N5, and N1, which completely conform to the evolutionary dynamics of influenza NAs (Xu et al., 2012) and the NJ trees drawn by ClustalW and MUSCLE.

B, and HRV-C, and HRV-A and HRV-C share a common ancestor, which is a sister group of HRV-B [32].

To clarify the classification of HRVs, a dataset containing 113 HRV and 3 HEV-C complete genomes is utilized to investigate the classification of HPVs, the lengths of which are between 2142 and 2214 amino acids. Comparing with results derived from genomic sequence, our result looks more credible, in that, protein sequences are more conservative in structure and function than nucleotide sequences [33]. As shown in Fig. 3, all 113 HRVs are clustered into three groups: HRV-A, HRV-B, and HRV-C, and 3 HEV-Cs form an outgroup, which are in accord with clinical heterogeneity of HRV infections in humans and results gotten by published methods [34–36].

The phylogenetic trees generated by ClustalW and MUSCLE are well classified in Fig. 4and Fig. 4s, and their topologies look very similar to that produced by our new method. Especially, the topological structure of HPV-As in our NJ tree is better than the genome tree for all known HRV serotypes based on the Maximum likelihood and Maximum parsimony [37]. For example, the hrv-46 arose by recombination between hrv-53 (major parent) and hrv-80 (minor parent), which are more obvious in our NJ tree.

3.3. Phylogenetic analysis of 1163 influenza A viruses

Influenza A viruses cause influenza in birds and some mammals, which are clinically significant in evoking several serious human illness [38]. Influenza A viruses are highly variable, and are classified according to the antigenic variation of surface glycoproteins: hemagglutinin (HA) and neuraminidase (NA) [39]. Therefore, the subtypes of influenza A viruses are named by an H number (for the type of hemagglutinin) and an N number (for the type of neuraminidase). To date, 18 antigenic variants of HA (from H1 to H18) and 11 antigenic variants of NA (from N1 to N11) have been recognized [40]. For example, H1N1, H2N2, H5N1, H7N3 and H7N9 are the most lethal subtypes in influenza A viruses [41].

A much larger dataset consisting of 1163 NA sequences that encode influenza A viruses is utilized to evaluate prediction accuracy in the classification of proteins [42,43], which are divided into 13 subtypes: H5N6, H5N1, H7N9, H1N1, H6N2, H3N8, H3N2, H4N6, H5N5, H10N3 and H7N3. The phylogenetic tree of 1163 NA sequences is shown in Fig. 5 by NJ method.

Looking at Fig. 5, all influenza A viruses are divided into two subgroups: one consisting the types of N2, N3, N6, and N9; and the other one containing the types of N8, N5, and N1. The viruses in each subgroup are independently adapted to different hosts, indicating that the parallel evolution occurs within the two subgroups due to the similar rates of genetic mutation and adaption to host environments. In addition, each of seven influenza A NA subtype forming a distinct cluster denotes a monophyletic origin for each subtype. The phylogenetic of N2-N3-N6-N9 and N8-N5-N1, which completely comply with the evolutionary dynamics of influenza NAs [44] and the NJ trees drawn by ClustalW and MUSCLE shown in Fig. 6 and Fig. 6s. Moreover, our NJ tree looks a little better than that of ClustalW, which is found from the cluster of EP1682990H1N1, EP1703429H1N1, and EP1239078 H5N1.

The computing efficiency is an important factor for all the new methods proposed for phylogenetic analysis of protein sequences that should perform sequence analysis within a limited time, even for whole genome data. To illustrate the efficiency of *k*-mer natural vector method, the computing time of our new method on beta-globin

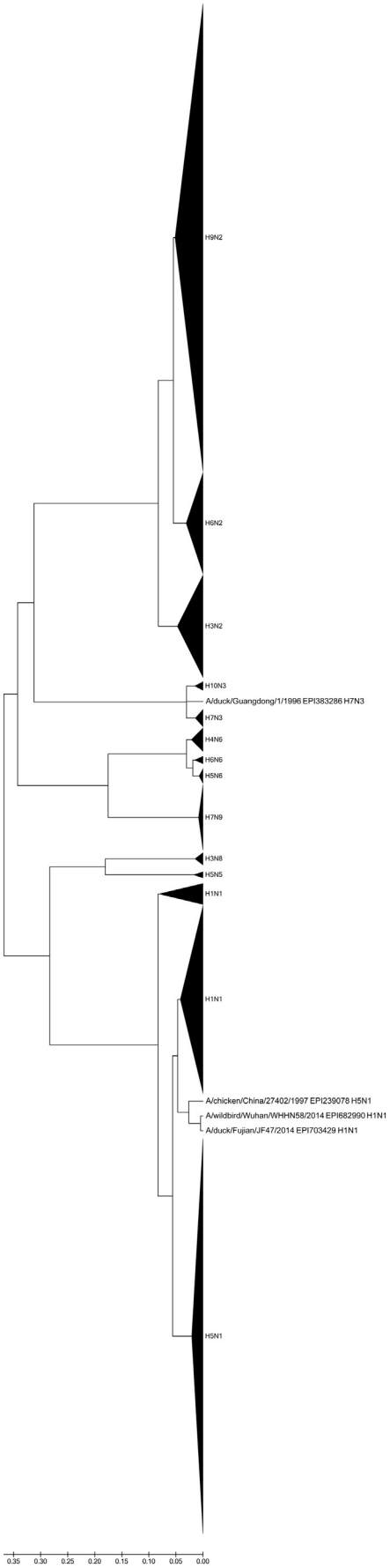

Fig. 6. NJ phylogenetic tree of 1163 influenza A viruses based on ClustalW. All 1163 influenza A viruses are divided into two subgroups: one consisting the types of N2, N3, N6, and N9; and the other one containing the types of N8, N5, and N1.

sequence, human rhinovirus, and influenza A virus are listed in Table 1, as well as ClustalW and MUSCLE. As shown in Table 1, our new method is more efficient than ClustalW and MUSCLE, which is easily found from Human rhinovirus and Influenza A virus, although MUSCLE runs a little faster on Beta-globin sequence.

In addition, for k-mer model method, the parameter k has a great influence on obtaining results and computational complexity. There exists no a criterion to choose a suitable k when different kinds of sequences are considered. Several methods have tried to find a suitable k. Huang and Yu utilized the stability of the distance matrix to find the optimal k [45]. Then, a k-string dictionary was proposed to use a lower dimensional frequency vector to represent a protein sequence [46]. Furthermore, the cross-validation was used to decide value of k in the k-nearest neighbor algorithm [47]. Since it is difficult to select a k that is feasible to all kinds of sequences, a compound k-mer natural vector with $k = 2$ and 3 is jointed to quantify each protein sequence, by which both long and short sequences can be accurately depicted.

4. Conclusions

Integrating the distributions of k-mers into k-mer model, a novel k-mer natural vector method is developed to accurately depict the evolutionary relationship of protein sequences, which contains the information on the relationships of k-mers to overcome the deficiency of former k-mer model methods. With this new method, the features of k-mers hidden in the sequence can be effectively extracted, and each protein sequence is numerically characterized by a compound k-mer natural vector. We mathematically prove that there exists a one-to-one relationship between a protein sequence and its associated k-mer natural vector for each given k. Therefore, phylogenetic analysis of protein sequences can be easily performed without requiring evolutionary models or human intervention.

We illustrate the utilities of this new method in exploring the phylogeny of protein sequences on the real data, by which our obtaining results are consistent with or better than the current most classic in phylogenetic analysis and published papers. We have verified that the k-mer natural vector method can not only improve the accuracies in depicting evolutionary relationship of protein sequences, but also strengthen the computing efficiency in dealing with more sequence data. Moreover, both long and short protein sequences can be effectively handled with our new method. However, our k-mer natural vector method is still in the process of being improved, it needs to remedy some disadvantages and drawbacks.

Conflict of interest statement

The authors declared no competing final interests.

Acknowledgements

We thank Prof. Craig Seeley and Prof. Changchuan Yin critically reading and editing our manuscript, and Dr. Lili He providing protein sequence data for our work. We also thank anonymous reviewers for their hard work and good suggestions. This work is supported by Youth Funding of Suihua University (K1501006), Scientific Research Funding of Suihua University (K1501009, 2017-XGYYWF-017), Scientific Research Funding of Heilongjiang Education Department (2017-KYYWF-0721). This study is also supported by the National Natural Science Foundation of China (91746119 to S. S.-T. Yau), Tsinghua University startup fund (to S. S.-T. Yau). Prof. S. S.-T. Yau is grateful to

Table 1
The computing time of *k*-mer natural vector method, ClustalW, and MUSCLE used on beta-globin sequence, human rhinovirus, and influenza A virus, respectively[a].

Dataset	*K*-mer natural vector method	ClustalW	MUSCLE
Beta-globin sequence	14.81 s	16.50 s	14.68 s
Human rhinovirus	32.78 s	38.40 min	1.14 min
Influenza A virus	4.01 min	4.00 h	6.40 min

[a] The configuration for our current laptop is Intel Core i5-2450 dual cores 2.50 GHZ with 8.00 Gb memory.

National Center for Theoretical Sciences (NCTS) for providing excellent research environment while part of this research was done.

Appendix A. Supplementary data

All datasets and matlab code used in this paper are available at https://github.com/wenjia198021/k-mer-natural-vector-for-protein-sequence.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2018.08.010.

References

[1] A. Rokas, Phylogenetic analysis of protein sequence data using the Randomized Axelerated Maximum Likelihood (RAXML) program, Current Protocols in Molecular Biology 96 (2011) 19.11.
[2] G. Dey, T. Meyer, Phylogenetic profiling for probing the modular architecture of the human genome, Cell Systems 1 (2015) 106–115.
[3] X.H. Xie, Z.G. Yu, G.S. Han, W.F. Yang, V. Anh, Whole-proteome based phylogenetic tree construction with inter-amino-acid distances and the conditional geometric distribution profiles, Molecular Phylogenetics and Evolution 89 (2015) 37–45.
[4] Y. Li, K. Tian, C. Yin, R.L. He, S.S.-T. Yau, Virus classification in 60-dimensional protein space, Molecular Phylogenetics and Evolution 99 (2016) 53–62.
[5] L. He, Y. Li, R.L. He, S.S.-T. Yau, A novel alignment-free vector method to cluster protein sequences, Journal of Theoretical Biology 427 (2017) 41–52.
[6] I.J. Kitching, P.L. Forey, D. Williams, C. Humphries, Cladistics: The Theory and Practice of Parsimony Analysis, Oxford University Press, USA, 1998.
[7] M. Nei, S. Kumar, Molecular Evolution and Phylogenetics, Oxford university press, 2000.
[8] C. Notredame, D.G. Higgins, J. Heringa, T-coffee: a novel method for fast and accurate multiple sequence alignment1, Journal of Molecular Biology 302 (2000) 205–217.
[9] J.P. Huelsenbeck, F. Ronquist, R. Nielsen, J.P. Bollback, Bayesian inference of phylogeny and its impact on evolutionary biology, Science 294 (2001) 2310–2314.
[10] K. Katoh, K. Misawa, K.I. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Research 30 (2002) 3059–3066.
[11] A. Löytynoja, N. Goldman, Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis, Science 320 (2008) 1632–1635.
[12] M.R. Kantorovitz, G.E. Robinson, S. Sinha, A statistical method for alignment-free comparison of regulatory sequences, Bioinformatics 23 (2007) i249–i255.
[13] T.J. Wu, Y.H. Huang, L.A. Li, Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences, Bioinformatics 21 (2005) 4125–4132.
[14] G.E. Sims, S.R. Jun, G.A. Wu, S.H. Kim, Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions, Proceedings of the National Academy of Sciences 106 (2009) 2677–2682.
[15] H.J. Yu, Segmented K-mer and its application on similarity analysis of mitochondrial genome sequences, Gene 518 (2013) 419–424.
[16] J. Wen, Y. Zhang, S.S.-T. Yau, K-mer Sparse matrix model for genetic sequence and its applications in sequence comparison, Journal of Theoretical Biology 363 (2014) 145–150.
[17] M. Deng, C. Yu, Q. Liang, R.L. He, S.S.-T. Yau, A novel method of characterizing genetic sequences: genome space with biological distance and applications, PLoS ONE 6 (2011) e17293.
[18] C. Yu, M. Deng, S.Y. Cheng, S.C. Yau, R.L. He, S.S.-T. Yau, Protein space: a natural method for realizing the nature of protein universe, Journal of Theoretical Biology 318 (2013) 197–204.
[19] J. Wen, R.H. Chan, S.C. Yau, R.L. He, S.S.-T. Yau, *K*-mer natural vector and its application to the phylogenetic analysis of genetic sequences, Gene 546 (2014) 25–34.
[20] M.W. Berry, Z. Drmac, E.R. Jessup, Matrices, vector spaces, and information retrieval, SIAM Review 41 (1999) 335–362.
[21] G.W. Stuart, K. Moffett, J.J. Leader, A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes, Molecular Biology and Evolution 19 (2002) 554–562.
[22] G.W. Stuart, M.W. Berry, An SVD-based comparison of nine whole eukaryotic genomes supports a coelomate rather than ecdysozoan lineage, BMC Bioinformatics 5 (2004) 204.
[23] J. Qi, B. Wang, B.I. Hao, Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach, Journal of Molecular Evolution 58 (2004) 1–11.
[24] K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, MEGA6: molecular evolutionary genetics analysis version 6.0, Molecular Biology and Evolution 30 (2013) 2725–2729.
[25] S.S.-T. Yau, C. Yu, R. He, A protein map and its application, DNA and Cell Biology 27 (2008) 241–250.
[26] T.J. Near, R.I. Eytan, A. Dornburg, K.L. Kuhn, J.A. Moore, M.P. Davis, P.C. Wainwright, M. Friedman, W.L. Smith, Resolution of ray-finned fish phylogeny and timing of diversification, Proceedings of the National Academy of Sciences 109 (2012) 13698–13703.
[27] C.G. Sibley, J.E. Ahlquist, B.L. Monroe Jr., A Classification of the Living Birds of the World Based on DNA-DNA Hybridization Studies. The Auk, (1988), pp. 409–423.
[28] A.L. Chubb, New nuclear evidence for the oldest divergence among neognath birds: the phylogenetic utility of ZENK (i), Molecular Phylogenetics and Evolution 30 (2004) 140–151.
[29] J.O. Kriegs, A. Matzke, G. Churakov, A. Kuritzin, G. Mayr, J. Brosius, J. Schmitz, Waves of genomic hitchhikers shed light on the evolution of gamebirds, BMC Evolutionary Biology 7 (2007) 190.
[30] A. Ruohola, M. Waris, T. Allander, T. Ziegler, T. Heikkinen, O. Ruuskanen, Viral etiology of common cold in children, Finland, Emerging Infectious Diseases 15 (2009) 344.
[31] A.C. Palmenberg, J.E. Gern, Classification and evolution of human rhinoviruses, Rhinoviruses, 1–10, Humana Press, New York, NY, 2015.
[32] S. Milanoi, J.R. Ongus, G. Gachara, R. Coldren, W. Bulimo, Serotype and genetic diversity of human rhinovirus strains that circulated in Kenya in 2008, Influenza and Other Respiratory Viruses 10 (2016) 185–191.
[33] F.R. Opperdoes, Phylogenetic Analysis Using Protein Sequences. The Phylogenetics Handbook A Practical Approach to DNA and Protein Phylogeny, (2003), pp. 207–235.
[34] J.M. Gwaltney Jr., J.O. Hendley, G. Simon, W.S. Jordan Jr., Rhinovirus infections in an industrial population: the occurrence of illness, The New England Journal of Medicine 275 (1966) 1261–1268.
[35] K.G. Nicholson, J. Kent, D.C. Ireland, Respiratory viruses and exacerbations of asthma in adults, BMJ 307 (1993) 982–986.
[36] D.J. Jackson, R.E. Gangnon, M.D. Evans, K.A. Roberg, E.L. Anderson, T.E. Pappas, M.C. Printz, W.M. Lee, P.A. Shult, E. Reisdorf, K.T. Carlson-Dakes, Wheezing rhinovirus illnesses in early life predict asthma development in high-risk children, American Journal of Respiratory and Critical Care Medicine 178 (2008) 667–672.
[37] A.C. Palmenberg, D. Spiro, R. Kuzmickas, S. Wang, A. Djikeng, J.A. Rathe, C.M. Fraser-Liggett, S.B. Liggett, Sequencing and analyses of all known human rhinovirus genomes reveal structure and evolution, Science 324 (2009) 55–59.
[38] M.W. Russell, J. Mestecky, W. Strober, B.N. Lambrecht, B.L. Kelsall, H. Cheroutre, Overview: the mucosal immune system, In Mucosal Immunology, 2015, pp. 3–8.
[39] R.G. Webster, W.J. Bean, O.T. Gorman, T.M. Chambers, Y. Kawaoka, Evolution and ecology of influenza A viruses, Microbiological Reviews 56 (1992) 152–179.
[40] K. Shinya, M. Ebina, S. Yamada, M. Ono, N. Kasai, Y. Kawaoka, Avian flu: influenza virus receptors in the human airway, Nature 440 (2006) 435.
[41] R.A. Fouchier, V. Munster, A. Wallensten, T.M. Bestebroer, S. Herfst, D. Smith, G.F. Rimmelzwaan, B. Olsen, A.D. Osterhaus, Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls, Journal of Virology 79 (2005) 2814–2822.
[42] H.H. Huang, C. Yu, H. Zheng, T. Hernandez, S.C. Yau, R.L. He, J. Yang, S.S.-T. Yau, Global comparison of multiple-segmented viruses in 12-dimensional genome space, Molecular Phylogenetics and Evolution 81 (2014) 29–36.
[43] T. Hoang, C. Yin, H. Zheng, C. Yu, R.L. He, S.S.-T. Yau, A new method to cluster DNA sequences using Fourier power spectrum, Journal of Theoretical Biology 372 (2015) 135–145.
[44] J. Xu, C.T. Davis, M.C. Christman, P. Rivailler, H. Zhong, R.O. Donis, G. Lu, Evolutionary history and phylodynamics of influenza A and B neuraminidase (NA) genes inferred from large-scale sequence analyses, PLoS ONE 7 (2012) e38665.
[45] H.H. Huang, C. Yu, Clustering DNA sequences using the out-of-place measure with reduced n-grams, Journal of Theoretical Biology 406 (2016) 61–72.
[46] C. Yu, R.L. He, S.S.-T. Yau, Protein sequence comparison based on K-string dictionary, Gene 529 (2013) 250–256.
[47] T. Hernandez, J. Yang, Descriptive statistics of the genome: phylogenetic classification of viruses, Journal of Computational Biology 23 (2016) 810–820.