Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

A novel alignment-free vector method to cluster protein sequences

Lily He[a,1], Yongkun Li[a,1], Rong Lucy He[b], Stephen S.-T. Yau[a,*]

[a] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, PR China
[b] Department of Biological Sciences, Chicago State University, Chicago, IL, USA

ARTICLE INFO

Article history:
Received 17 February 2017
Revised 4 May 2017
Accepted 2 June 2017
Available online 3 June 2017

Keywords:
Phylogeny
Biochemical properties
Vector
Alignment-free

ABSTRACT

Classification of protein are crucial topics in biology. The number of protein sequences stored in databases increases sharply in the past decade. Traditionally, comparison of protein sequences is usually carried out through multiple sequence alignment methods. However, these methods may be unsuitable for clustering of protein sequences when gene rearrangements occur such as in viral genomes. The computation is also very time-consuming for large datasets with long genomes. In this paper, based on three important biochemical properties of amino acids: the hydropathy index, polar requirement and chemical composition of the side chain, we propose a 24 dimensional feature vector describing the composition of amino acids in protein sequences. Our method not only utilizes the chemical properties of amino acids but also counts on their numbers and positions. The results on beta-globin, mammals, and three virus datasets show that this new tool is fast and accurate for classifying proteins and inferring the phylogeny of organisms.

Published by Elsevier Ltd.

1. Introduction

With the rapid increase of genetic data, a growing number of methods for sequence comparison have been proposed. Most of them are alignment-based. These methods may cluster gene sequence precisely and several algorithms have been accepted widely (Edgar, 2004; Katoh et al., 2002). Nevertheless, these methods often consume long run time and cause heavy burden of memory. In addition, these methods may become limited due to the high mutation rate and recombination as in viral genomes. Alignment-free approaches have already sprung up and been successfully applied in biological fields (Deng et al., 2011; Li et al., 2017; Vinga and Almeida, 2003; Yin et al., 2014). Specifically, K-mer cluster method was proposed and used widely (Blaisdell, 1989). The k-mer words in a protein sequence are substrings with fixed length k. There are 20^k all possible kinds of k-mer words in protein sequences. The k-mer method assembles the frequencies of all possible k-mer words into a numerical vector whose dimension is 20^k. However, k-mer method only considers the sequence context without using chemical properties and positions of amino acids.

Understanding protein similarity relationships is vital for the annotation of genome sequences (Gan et al., 2002; Pearl et al., 2000). Proteins with high sequence identity tend to possess similarity in functions and evolutionary relationships. Therefore, using proteins to analyze the similarity of species makes more sense than using DNA sequences (Li et al., 2016; Xie et al., 2015). Although the classic k-mer model and their variants are widely used in many biological studies, the dimension of the numerical vector derived from a protein sequence is very high when k is large. (Jun et al., 2010; Qi et al., 2004; Ulitsky et al., 2006). For instance, when $k = 5$ the dimension of the k-mer vector is $20^5 = 3,200,000$. In this case, the computational complexity for counting the number of k-mer strings becomes high. Furthermore, the choice of k is difficult and depends on divergence of protein sequences. The position information and the chemical properties of amino acids are not considered in k-mer methods as well.

With the rapid growth of biological sequence, computational methods to analyze the data focus on development of suitable vector models to characterize related biological sequences, because all existing operation algorithms or engines are unable to use sequences directly (Chou, 2015). To avoid complete loss of sequence pattern, the pseudo amino acid composition (PseAAC) was proposed by Chou (2001). Since then, this method was applied to identify nucleosomal sequences, interactions of proteins and protein-protein binding sites (Chen et al., 2012; Jia et al., 2015; 2016b; Wu et al., 2010).The approach also achieved much successes in drug development areas (Zhong and Zhou, 2014) and many computational proteomics (Chou, 2009; Chou and Zhang, 1995; Jiao and Du, 2017; Khan et al., 2017; Kumar et al., 2015; Lin and Lapointe, 2013; Meher et al., 2017; Mondal and Pai, 2014). Since the wide application of PseAAC, some open access soft-wares were developed. The ‘PseAAC-Builder’ (Du et al., 2012), ‘propy’

[*] Corresponding author.
E-mail address: yau@uic.edu (S.S.-T. Yau).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jtbi.2017.06.002
0022-5193/Published by Elsevier Ltd.

Table 1
Three physicochemical properties of 20 amino acids.

Amino acids	Hydropathy index	Polar requirement	Chemical composition of the side chain
A(Ala)	1.8	7.0	0
C(Cys)	2.5	4.8	2.75
D(Asp)	−3.5	13.0	1.38
E(Glu)	−3.5	12.5	0.92
F(Phe)	2.8	5.0	0
G(Gly)	−0.4	7.9	0.74
H(His)	−3.2	8.4	0.58
I(Ile)	4.5	4.9	0
K(Lys)	−3.9	10.1	0.33
L(Leu)	3.8	4.9	0
M(Met)	1.9	5.3	0
N(Asn)	−3.5	10.0	1.33
P(Pro)	−1.6	6.6	0.39
Q(Gln)	−3.5	8.6	0.89
R(Arg)	−4.5	9.1	0.65
S(Ser)	−0.8	7.5	1.42
T(Thr)	−0.7	6.6	0.71
V(Val)	4.2	5.6	0
W(Trp)	−0.9	5.2	0.13
Y(Tyr)	−1.3	5.4	0.20

(Cao et al., 2013) may generate various modes of Chou's special PseAAC. The 'PseAAC-General' (Du et al., 2014) was aimed at producing various modes of Chou's general PseAAC which contains higher level feature vectors such as 'Functional Domain' mode, 'Gene Ontology' mode, and 'Sequential Evolution' (Chou, 2011). Three web-servers were also established for generating various feature vectors for biological sequences (Chen et al., 2014; 2015; Liu et al., 2015a). Recently a powerful web-server called Pse-in-One (Liu et al., 2015b) was designed to generate any desired feature vectors for sequences.

The physicochemical properties of amino acids are important for protein sequence classification and evolution (Salichos and Rokas, 2013; Wimley and White, 1996). There are more than ten kinds of properties well established (Rackovsky, 2009). Among them, the hydropathy index measures the hydrophilicity or hydrophobicity of amino acids. As the value of ***hydropathy index*** increases, an amino acid becomes more hydrophobic (Kyte and Doolittle, 1982; Yau et al., 2008). This factor plays important rules in prediction of protein structures and phylogenetic analysis. The ***polar requirement property*** represents the polarity of an amino acid, which is also crucial for protein studies (Woese et al., 1966). The third factor of an amino acid is the ***chemical composition of the side chain***, which affects some chemical properties globally (Grantham, 1974). The three properties are commonly used and have strong effects on protein structure and function.

In this paper, we find that use of the three physicochemical properties is helpful for the classification and evolution of protein sequences. For each property, the 20 amino acids may be grouped into several classes according to the value of this property. Then we propose a set of numbers to describe distribution of amino acids in each class in a protein sequence. Finally, we establish a novel 24 dimensional numerical vector to characterize each protein sequence. This vector takes into account both the positions of amino acids and the chemical properties of amino acids in protein sequences. Another advantage is that our vector method can be used by most of clustering algorithms which can not deal with sequences directly. To test the superiority of our novel method, we analyze several data sets and further compare our method with the popular alignment ClustalW algorithm. Our new tool is more accurate to infer phylogenetic relationships of organisms and much lower in computational complexity than ClustalW.

Table 2
Classification of the 20 amino acids.

	Amino acids	Denote
Hydropathy index		
> 0	A C F I L M V	α
$(-1, 0)$	G S T W	β
< -1	D E H K N P Q R Y	γ
Polar requirement		
≥ 7	A D E G H K N Q R S	δ
< 7	C F I L M P T V W Y	ϕ
Chemical composition of the side chain		
0	A F I L M V	ρ
(0, 1)	E G H K P Q R T W Y	θ
> 1	C D N S	λ

2. Methods

Let $\mathcal{F}$ be the set of 20 amino acids, i.e., $\mathcal{F} = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ and $S = (s_1, s_2, \cdots, s_N)$ be a protein sequence of length n, that is, $s_i \in \mathcal{F}, i = 1, 2, \cdots, N$. Here we consider 3 major physicochemical properties of amino acids including the ***hydropathy index*** (Kyte and Doolittle, 1982), the ***polar requirement*** (Woese et al., 1966) and the ***chemical composition of the side chain*** (Grantham, 1974). The values of these properties for amino acids are listed in Table 1. For the hydropathy index factor, the more positive the value is, the more hydrophobic an amino acid is. Amino acids with close values tend to have similar hydrophobicity. The positive value of an amino acid means it has hydrophobicity, while the negative value means it has hydrophilicity. So we split the 20 amino acids into 7 hydrophobic amino acids and 13 hydrophilic amino acids. Among the 13 hydrophilic amino acids, the hydropathy indexes of some are close to zero and others are relatively large. Therefore, we group the 13 amino acids into two subgroups further. Finally, we classify the twenty amino acids into 3 groups: $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3$, where: $\mathcal{H}_1 = \{A, C, F, I, L, M, V\}$; $\mathcal{H}_2 = \{G, S, T, W\}$; $\mathcal{H}_3 = \{D, E, H, K, N, P, Q, R, Y\}$. The hydropathy values of amino acids in $\mathcal{H}_1$ are greater than 0. The hydropathy values of amino acids from $\mathcal{H}_2$ are in the interval $(-1, 0)$. For amino acids from $\mathcal{H}_3$, their hydropathy values are less than -1. We denote the three groups by α, β and γ respectively Table 2.

For the polar requirement property, the values for the 20 amino acids seem even. So we divide the amino acids into two cate-

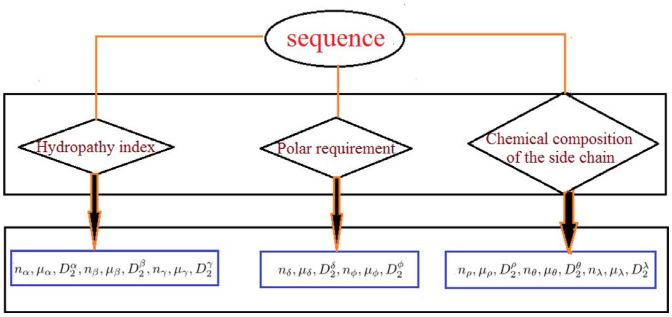
sequence
Hydropathy index
Polar requirement
Chemical composition of the side chain
$n_\alpha, \mu_\alpha, D_2^\alpha, n_\beta, \mu_\beta, D_2^\beta, n_\gamma, \mu_\gamma, D_2^\gamma$
$n_\delta, \mu_\delta, D_2^\delta, n_\phi, \mu_\phi, D_2^\phi$
$n_\rho, \mu_\rho, D_2^\rho, n_\theta, \mu_\theta, D_2^\theta, n_\lambda, \mu_\lambda, D_2^\lambda$

Fig. 1. Flow chart of the construction of our new 24 dimensional vector. The 20 amino acids are divided into tree groups α, β and γ according to the hydropathy index, δ and ϕ according to the polar requirement, ρ, θ and λ according to the chemical composition of the side chain. For each group, three numerical features are proposed to describe count, center and variability of the group in the protein sequence.

gories with equal size : $\mathcal{P}_1 = \{A,\ D,\ E,\ G,\ H,\ K,\ N,\ Q,\ R,\ S\}$ and $\mathcal{P}_2 = \{C,\ F,\ I,\ L,\ M,\ P,\ T,\ V,\ W,\ Y\}$. We denote the amino acids in $\mathcal{P}_1$ and $\mathcal{P}_2$ by δ and ϕ respectively. The polar requirement values for amino acids in $\mathcal{P}_1$ are ≥ 7. The values for remaining 10 amino acids in $\mathcal{P}_2$ are < 7 as shown in Table 2. For the chemical composition property, the values of 6 amino acids are zero. We classify the 6 amino acids into a group. For the remaining 14 amino acids, a few of them seem different due to large value for this property. Thus we cluster the 20 amino acids into three classes by chemical composition of the side chain. The three classes are $\mathcal{C}_1 = \{A,\ F,\ I,\ L,\ M,\ V\}$, $\mathcal{C}_2 = \{E,\ G,\ H,\ K,\ P,\ Q,\ R,\ T,\ W,\ Y\}$, and $\mathcal{C}_3 = \{C,\ D,\ N,\ S\}$. We record the amino acids from the three classes as ρ, θ and λ respectively. The chemical composition of the side chain of amino acids in $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ are 0 , (0, 1), and > 1 respectively in Table 2.

As shown in the flow chart (Fig. 1), we first classify amino acids into three groups based on the hydropathy index. Then a protein sequence only contains three kinds of letters: α, β and γ. For α, we define $f_\alpha(\ \cdot\ )$: $\{\alpha, \beta, \gamma\} \rightarrow \{0,\ 1\}$, such that:

$$f_\alpha(s_i) = \begin{cases} 1, & s_i = \alpha \\ 0, & s_i \neq \alpha \end{cases} \qquad i = 1, 2, \cdots, N.$$

For α, we propose three features n_α, μ_α and D_2^α to describe the number of α, the average position of α and the variation of the position of α appearing in the sequence S. These features are defined as follows:

$$n_\alpha = \sum_{i=1}^{N} f_\alpha(s_i); \quad \mu_\alpha = \sum_{i=1}^{N} i \cdot \frac{f_\alpha(s_i)}{n_\alpha}; \quad D_2^\alpha = \sum_{i=1}^{N} \frac{(i - \mu_\alpha)^2 f_\alpha(s_i)}{n_\alpha \cdot N}.$$

For the other letters: $\beta, \gamma, \delta, \phi, \rho, \theta, \lambda$, we defined $f_\beta(\cdot)$,$f_\gamma(\cdot)$,$f_\delta(\cdot)$,$f_\phi(\cdot)$, $f_\rho(\cdot)$,$f_\theta(\cdot)$,$f_\lambda(\cdot)$ in the same way. For every letter we thus gain 3 features to describe its distribution in a protein sequence. These features form a 24 dimensional vector defined as

$$(n_\alpha,\ \mu_\alpha,\ D_2^\alpha,\ n_\beta,\ \mu_\beta,\ D_2^\beta,\ n_\gamma,\ \mu_\gamma,\ D_2^\gamma,\ n_\delta,\ \mu_\delta,\ D_2^\delta,\ n_\phi,\ \mu_\phi,\ D_2^\phi,\ n_\rho,\ \mu_\rho,\ D_2^\rho,\ n_\theta,\ \mu_\theta, D_2^\theta,\ n_\lambda,\ \mu_\lambda,\ D_2^\lambda).$$

Then the Euclidean distance is applied to calculate the pairwise distance among the 24 dimensional vectors of protein sequences. The phylogenetic tree of organisms can be built by using the UPGMA algorithm based on MEGA 7.0 software (Tamura et al., 2013).

3. Results

3.1. beta-globin

In order to verify our method, firstly we apply our method to classification of Beta-globin proteins which are the most common haemoglobin in adult humans (Yu et al., 2013). 50 beta-globin sequences picked from Swiss-Prot were analyzed using the protein map method by (Yau et al., 2008). In this study, 88 beta-globin sequences from more diverse species were extracted from Swiss-Prot (http://www.uniprot.org/). Using our method, these 88 sequences are clustered correctly into 19 groups: *Primates, Proboscidea, Carnivora, Hyracoidea, Insectivora, Columbiformes, Perissodactyla, Testudines, Salmoniformes, Cypriniformes, Diprotodontia, Galliformes, Passeriformes, Anseriformes, Sirenia, Rodentia, Anura, Gadiformes* and *Perciformes* Fig. 2. In contrast, the phylogenetic tree of these proteins are constructed by ClustalW algorithm. However, three protein sequences (marked in red) are misplaced as shown in Fig. 3.

3.2. Human rhinovirus

Human rhinovirus (HRV) belongs to genus *Enterovirus* and family *Picornaviridae*. Past studies have classified HRV into three genetically distinct groups, HRV-A, HRV-B, and HRV-C, within the genus *Enterovirus* (Deng et al., 2011; Hoang et al., 2015). From (Jacobs et al., 2013; Palmenberg et al., 2009) HRV-A and HRV-C share a common ancestor, which is a sister group to the HRV-B. Our results based on the new feature vector method are consistent with theirs Fig. 4. Note that these papers all use the complete genome sequences, but our paper use the concatenated protein sequences which are also polyproteins. In previous work (Palmenberg et al., 2009), a dataset consisting of 113 HRV and 3 HEV-C complete genomes was investigated. The 113 HRV genomes were clustered into three groups HRV-A, HRV-B, HRV-C and 3 HEV-C sequences formed an outgroup. While the genomes were well classified, the running time was quite high due to usage of multiple sequence alignment for constructing the evolutionary tree. In this paper, 114 polyprotein sequences are studied since 2 genomes with problematic polyproteins are excluded. The phylogenetic tree based on the new method is shown in Fig. 4. The running only takes 0.77 s to finish the conversion from sequences to numerical vectors in our laptop. Moreover, a phylogenetic tree is also produced by ClustalW and it takes 37 min to complete the multiple sequence alignment. The topology of the alignment-based tree is totally same as that by our new method.

3.3. Influenza A viruses

Influenza spreads around the world in a yearly outbreak, resulting in about three to five million cases of severe illness and about 250,000 to 500,000 deaths (de Jong et al., 2005). Three types of influenza viruses, Type A, Type B, and Type C, may cause damage to human health. Influenza A virus causes influenza in birds and some mammals. Influenza A viruses are negative-sense, single-stranded, segmented RNA viruses. The several subtypes are labeled according to an H number (for the type of hemagglutinin) and an N number (for the type of neuraminidase). There are 18 different known H antigens (H1 to H18) and 11 different known N antigens (N1 to N11). For example, H17 was isolated from fruit bats and H18N11 was discovered in a Peruvian bat in Tong et al. (2013). Influenza A viruses have caused many pandemics and some of the most lethal subtypes are H1N1, H2N2, H5N1, H7N3, and H7N9. These subtypes are chosen to test the efficiency of our method. Specifically, we examine 35 neuraminidase (NA) sequences encoded by the neuraminidase (NA) gene of Influenza

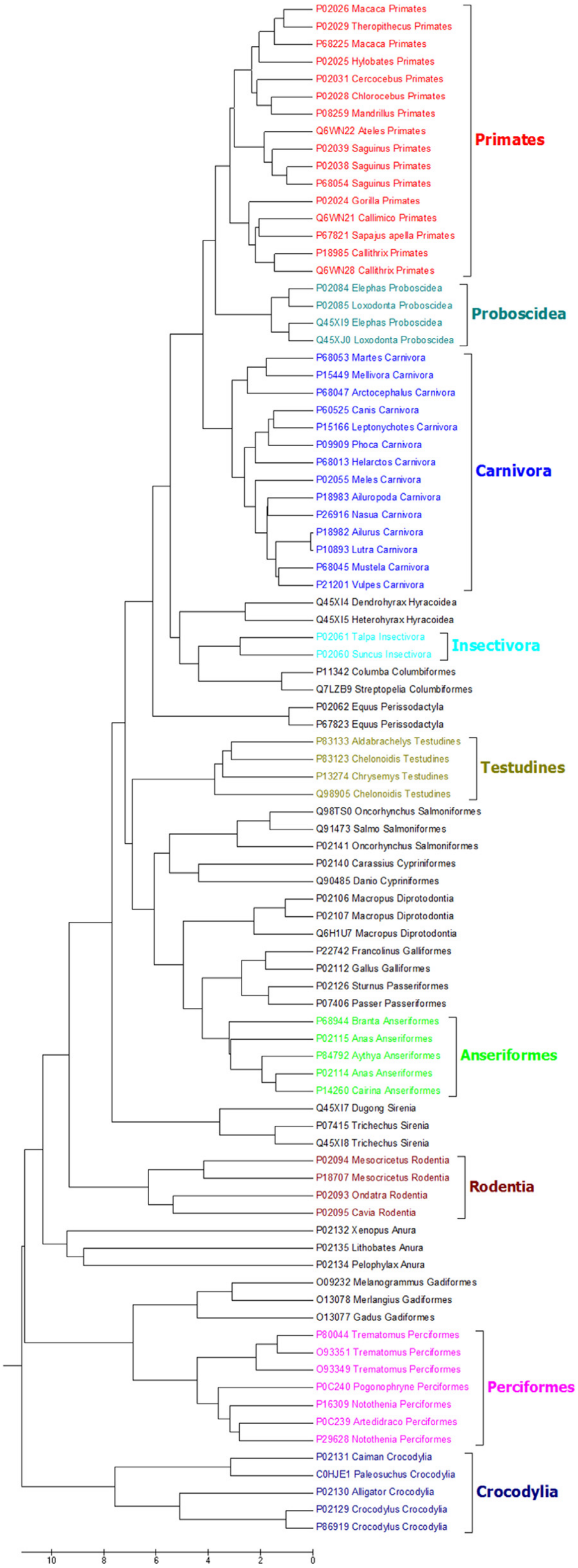
P02026 Macaca Primates
P02029 Theropithecus Primates
P68225 Macaca Primates
P02025 Hylobates Primates
P02031 Cercocebus Primates
P02028 Chlorocebus Primates
P08259 Mandrillus Primates
Q6WN22 Ateles Primates
P02039 Saguinus Primates
P02038 Saguinus Primates
P68054 Saguinus Primates
P02024 Gorilla Primates
Q6WN21 Callimico Primates
P67821 Sapajus apella Primates
P18985 Callithrix Primates
Q6WN28 Callithrix Primates
Primates
P02084 Elephas Proboscidea
P02085 Loxodonta Proboscidea
Q45XI9 Elephas Proboscidea
Q45XJ0 Loxodonta Proboscidea
Proboscidea
P68053 Martes Carnivora
P15449 Mellivora Carnivora
P68047 Arctocephalus Carnivora
P60525 Canis Carnivora
P15166 Leptonychotes Carnivora
P09909 Phoca Carnivora
P68013 Helarctos Carnivora
P02055 Meles Carnivora
P18983 Ailuropoda Carnivora
P26916 Nasua Carnivora
P18982 Ailurus Carnivora
P10893 Lutra Carnivora
P68045 Mustela Carnivora
P21201 Vulpes Carnivora
Carnivora
Q45XI4 Dendrohyrax Hyracoidea
Q45XI5 Heterohyrax Hyracoidea
P02061 Talpa Insectivora
P02060 Suncus Insectivora
Insectivora
P11342 Columba Columbiformes
Q7LZB9 Streptopelia Columbiformes
P02062 Equus Perissodactyla
P67823 Equus Perissodactyla
P83133 Aldabrachelys Testudines
P83123 Chelonoidis Testudines
P13274 Chrysemys Testudines
Q98905 Chelonoidis Testudines
Testudines
Q98TS0 Oncorhynchus Salmoniformes
Q91473 Salmo Salmoniformes
P02141 Oncorhynchus Salmoniformes
P02140 Carassius Cypriniformes
Q90485 Danio Cypriniformes
P02106 Macropus Diprotodontia
P02107 Macropus Diprotodontia
Q6H1U7 Macropus Diprotodontia
P22742 Francolinus Galliformes
P02112 Gallus Galliformes
P02126 Sturnus Passeriformes
P07406 Passer Passeriformes
P68944 Branta Anseriformes
P02115 Anas Anseriformes
P84792 Aythya Anseriformes
P02114 Anas Anseriformes
P14260 Cairina Anseriformes
Anseriformes
Q45XI7 Dugong Sirenia
P07415 Trichechus Sirenia
Q45XI8 Trichechus Sirenia
P02094 Mesocricetus Rodentia
P18707 Mesocricetus Rodentia
P02093 Ondatra Rodentia
P02095 Cavia Rodentia
Rodentia
P02132 Xenopus Anura
P02135 Lithobates Anura
P02134 Pelophylax Anura
O09232 Melanogrammus Gadiformes
O13078 Merlangius Gadiformes
O13077 Gadus Gadiformes
P80044 Trematomus Perciformes
O93351 Trematomus Perciformes
O93349 Trematomus Perciformes
P0C240 Pogonophryne Perciformes
P16309 Notothenia Perciformes
P0C239 Artedidraco Perciformes
P29628 Notothenia Perciformes
Perciformes
P02131 Caiman Crocodylia
C0HJE1 Paleosuchus Crocodylia
P02130 Alligator Crocodylia
P02129 Crocodylus Crocodylia
P86919 Crocodylus Crocodylia
Crocodylia
10
8
6
4
2
0

Fig. 2. UPGMA phylogenetic tree of beta-globin protein sequences of 88 species based on our new method. The dataset includes 20 groups: *Primates* (red), *Carnivora* (blue), *Proboscidea* (cyan), *Rodentia* (maroon), *Anseriformes* (green), *Insectivora* (aqua), *Perciformes* (purplish red), *Crocodylia* (navy blue), *Testudines* (olive), *Perissodactyla, Sirenia, Hyracoidea, Diprotodontia, Columbiformes, Passeriformes, Galliformes, Anura, Salmoniformes, Cypriniformes, Gadiformes*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

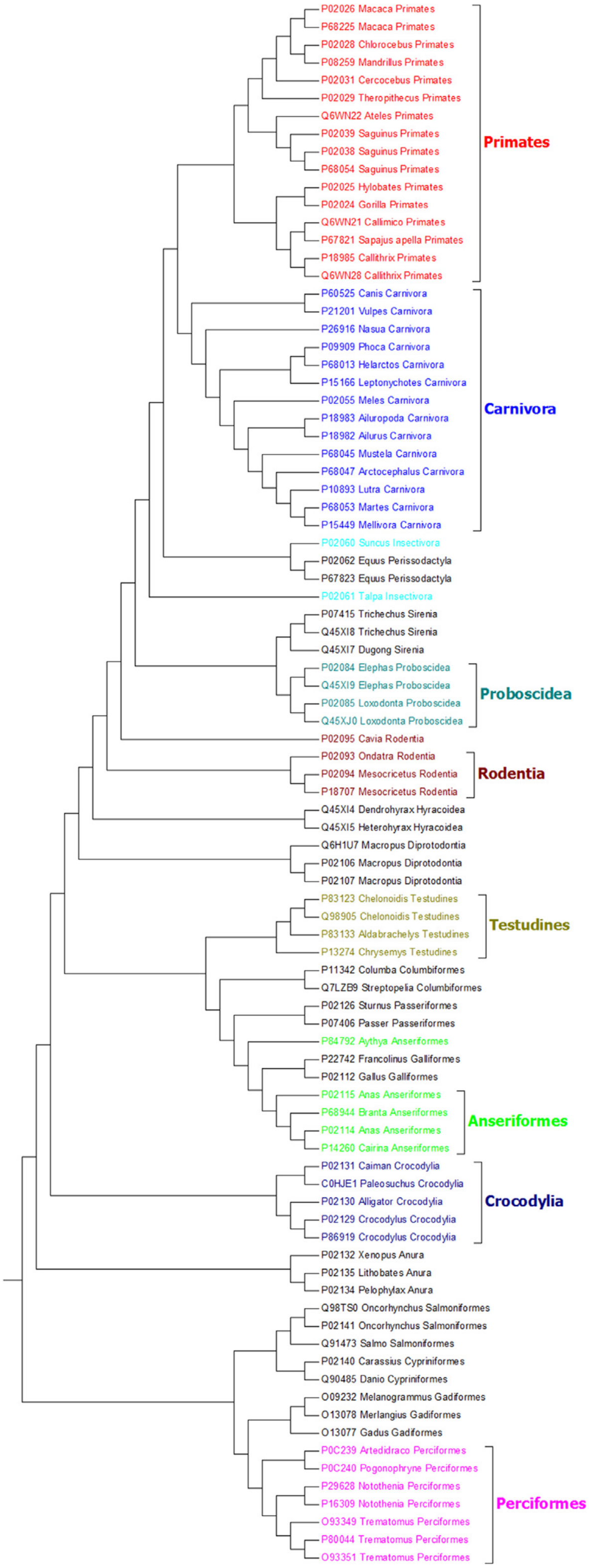
P02026 Macaca Primates
P68225 Macaca Primates
P02028 Chlorocebus Primates
P08259 Mandrillus Primates
P02031 Cercocebus Primates
P02029 Theropithecus Primates
Q6WN22 Ateles Primates
P02039 Saguinus Primates
P02038 Saguinus Primates
P68054 Saguinus Primates
P02025 Hylobates Primates
P02024 Gorilla Primates
Q6WN21 Callimico Primates
P67821 Sapajus apella Primates
P18985 Callithrix Primates
Q6WN28 Callithrix Primates
Primates
P60525 Canis Carnivora
P21201 Vulpes Carnivora
P26916 Nasua Carnivora
P09909 Phoca Carnivora
P68013 Helarctos Carnivora
P15166 Leptonychotes Carnivora
P02055 Meles Carnivora
P18983 Ailuropoda Carnivora
P18982 Ailurus Carnivora
P68045 Mustela Carnivora
P68047 Arctocephalus Carnivora
P10893 Lutra Carnivora
P68053 Martes Carnivora
P15449 Mellivora Carnivora
Carnivora
P02060 Suncus Insectivora
P02062 Equus Perissodactyla
P67823 Equus Perissodactyla
P02061 Talpa Insectivora
P07415 Trichechus Sirenia
Q45XI8 Trichechus Sirenia
Q45XI7 Dugong Sirenia
P02084 Elephas Proboscidea
Q45XI9 Elephas Proboscidea
P02085 Loxodonta Proboscidea
Q45XJ0 Loxodonta Proboscidea
Proboscidea
P02095 Cavia Rodentia
P02093 Ondatra Rodentia
P02094 Mesocricetus Rodentia
P18707 Mesocricetus Rodentia
Rodentia
Q45XI4 Dendrohyrax Hyracoidea
Q45XI5 Heterohyrax Hyracoidea
Q6H1U7 Macropus Diprotodontia
P02106 Macropus Diprotodontia
P02107 Macropus Diprotodontia
P83123 Chelonoidis Testudines
Q98905 Chelonoidis Testudines
P83133 Aldabrachelys Testudines
P13274 Chrysemys Testudines
Testudines
P11342 Columba Columbiformes
Q7LZB9 Streptopelia Columbiformes
P02126 Sturnus Passeriformes
P07406 Passer Passeriformes
P84792 Aythya Anseriformes
P22742 Francolinus Galliformes
P02112 Gallus Galliformes
P02115 Anas Anseriformes
P68944 Branta Anseriformes
P02114 Anas Anseriformes
P14260 Cairina Anseriformes
Anseriformes
P02131 Caiman Crocodylia
C0HJE1 Paleosuchus Crocodylia
P02130 Alligator Crocodylia
P02129 Crocodylus Crocodylia
P86919 Crocodylus Crocodylia
Crocodylia
P02132 Xenopus Anura
P02135 Lithobates Anura
P02134 Pelophylax Anura
Q98TS0 Oncorhynchus Salmoniformes
P02141 Oncorhynchus Salmoniformes
Q91473 Salmo Salmoniformes
P02140 Carassius Cypriniformes
Q90485 Danio Cypriniformes
O09232 Melanogrammus Gadiformes
O13078 Merlangius Gadiformes
O13077 Gadus Gadiformes
P0C239 Artedidraco Perciformes
P0C240 Pogonophryne Perciformes
P29628 Notothenia Perciformes
P16309 Notothenia Perciformes
O93349 Trematomus Perciformes
P80044 Trematomus Perciformes
O93351 Trematomus Perciformes
Perciformes

Fig. 3. UPGMA phylogenetic tree of beta-globin protein sequences of 88 species based on ClustalW. The dataset includes 20 groups: *Primates* (red), *Carnivora* (blue), *Proboscidea* (cyan), *Rodentia* (maroon), *Anseriformes* (green), *Insectivora* (aqua), *Perciformes* (purplish red), *Crocodylia* (navy blue), *Testudines* (olive), *Perissodactyla, Sirenia, Hyracoidea, Diprotodontia, Columbiformes, Passeriformes, Galliformes, Anura, Salmoniformes, Cypriniformes, Gadiformes*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

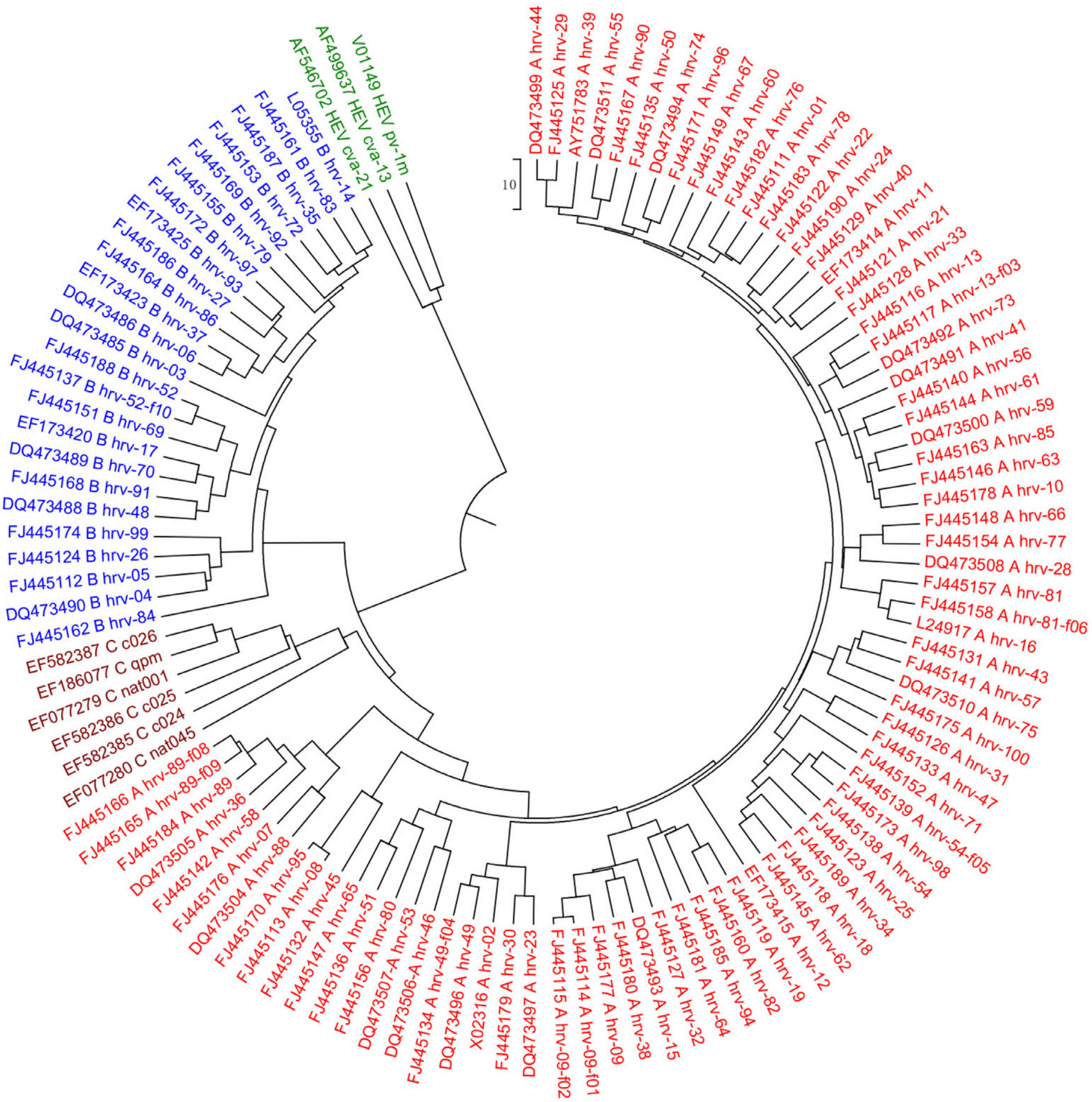

Fig. 4. UPGMA phylogenetic tree of 114 HRV serotypes using our method based on protein sequences. The HEV-C sequences (poliovirus 1M, coxsackievirus a13, and coxsackievirus a21) are used as outgroup. The dataset includes 4 groups: HRV-A (red), HRV-B (blue), HRV-C (maroon), HEV-C (green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

A virus (Webster et al., 1992). This dataset was analyzed by Fourier power spectrum approach using DNA sequences (Hoang et al., 2015; Huang et al., 2014). Based on our new method, the phylogenetic tree of 35 proteins are constructed. As illustrated in Fig. 5, the Influenza A viruses are clustered correctly. As comparison, the species are not clustered well in the tree constructed by ClustalW. As illustrated Fig. 6, A/turkey/VA/505477-18/2007(H5N1)), A/turkey/Ontario/FAV110-4/2009 (H1N1) and A/turkey/Virginial/4135/2014 (H1N1) are not correctly clustered. This may be caused by gene rearrangements occurring in the Influenza A genomes.

3.4. Mammalian mitochondria

The initial characterization of the mitochondrial proteome represents perhaps an even more important milestone for mitochondrial biology and medicine. Thus it is suitable to infer molecular evolution of mammals. In previous study (Tobe et al., 2010), the cytochrome b gene (cyt b) was shown to be very accurate to reconstruct mammalian phylogeny at super order, order ,family and generic levels. The protein of cyt b gene has about 380 amino acids. In our method, 79 mammalian cyte b protein sequences from the study are chosen. These sequences are clustered correctly into 10 groups: Carnivore, Chiroptera, Soricomorpha, Rodentia, Perissodactyla, Artiodactyla, Dipro- todontia, Monotremata, Cetacea and Primates Fig. 7. In contrast, evolutionary tree constructed by ClustalW is built as well. As illustrated in Fig. 8, several mammals (in red) are placed in the wrong location.

3.5. Prediction accuracy rate of current method

In order to test our new approach, we evaluate prediction accuracy rates by cross validation using leave-one-out method and first nearest neighbor (1-NN) algorithm. The results for above four datasets are shown in Table 3. The prediction accuracy rates for the datasets obtained by alignment algorithm ClustalW are also listed in this table. As we can see in Table 3, our method achieves higher accuracy rates than ClustalW for the Beta-globin data, 35 Influenza A viruses data (Influenza dataset1). For the HRV dataset with 114 viruses, the two methods both achieve 100% accuracy rate. For the

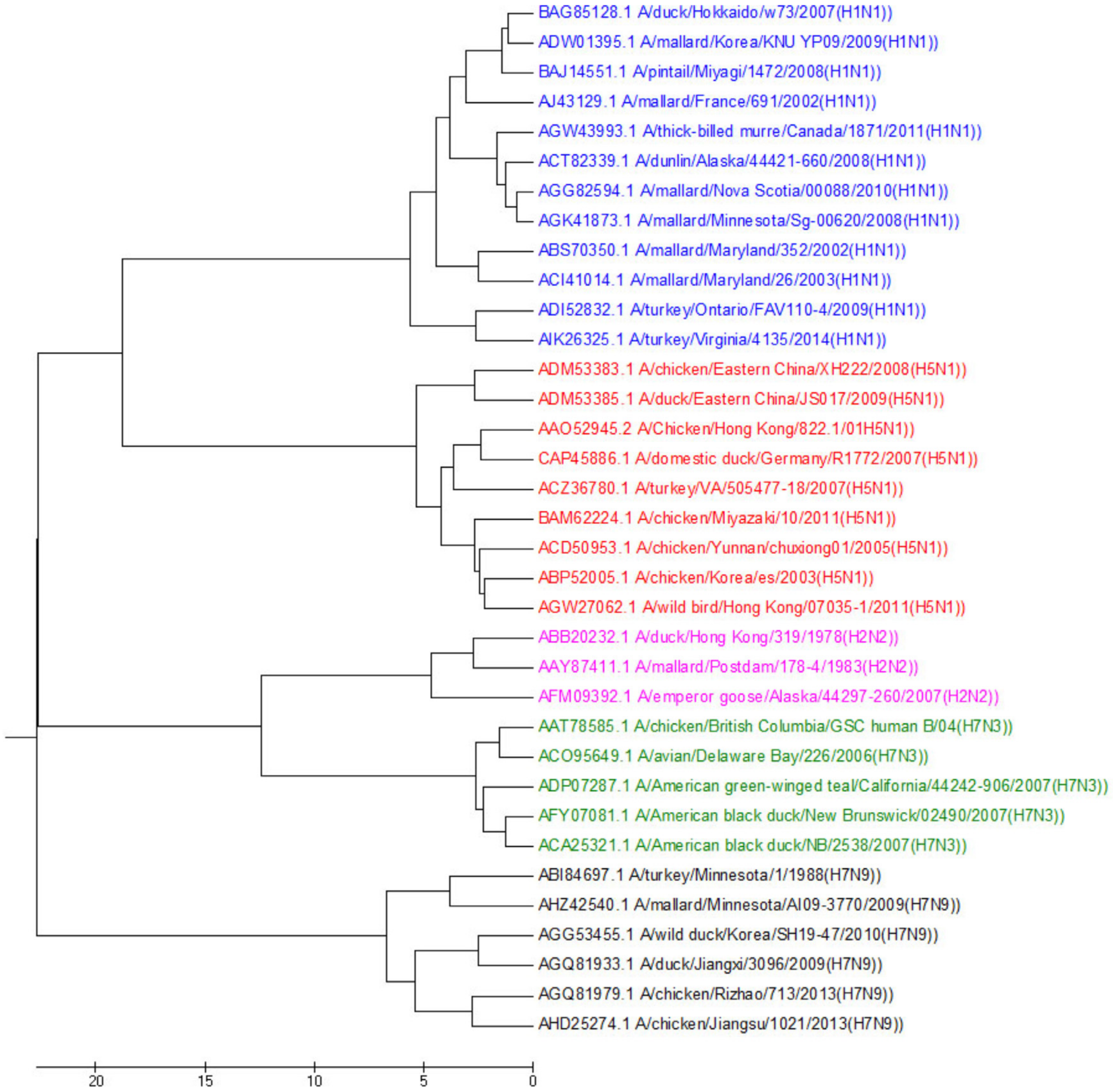
BAG85128.1 A/duck/Hokkaido/w73/2007(H1N1))
ADW01395.1 A/mallard/Korea/KNU YP09/2009(H1N1))
BAJ14551.1 A/pintail/Miyagi/1472/2008(H1N1))
AJ43129.1 A/mallard/France/691/2002(H1N1))
AGW43993.1 A/thick-billed murre/Canada/1871/2011(H1N1))
ACT82339.1 A/dunlin/Alaska/44421-660/2008(H1N1))
AGG82594.1 A/mallard/Nova Scotia/00088/2010(H1N1))
AGK41873.1 A/mallard/Minnesota/Sg-00620/2008(H1N1))
ABS70350.1 A/mallard/Maryland/352/2002(H1N1))
ACI41014.1 A/mallard/Maryland/26/2003(H1N1))
ADI52832.1 A/turkey/Ontario/FAV110-4/2009(H1N1))
AIK26325.1 A/turkey/Virginia/4135/2014(H1N1))
ADM53383.1 A/chicken/Eastern China/XH222/2008(H5N1))
ADM53385.1 A/duck/Eastern China/JS017/2009(H5N1))
AAO52945.2 A/Chicken/Hong Kong/822.1/01H5N1))
CAP45886.1 A/domestic duck/Germany/R1772/2007(H5N1))
ACZ36780.1 A/turkey/VA/505477-18/2007(H5N1))
BAM62224.1 A/chicken/Miyazaki/10/2011(H5N1))
ACD50953.1 A/chicken/Yunnan/chuxiong01/2005(H5N1))
ABP52005.1 A/chicken/Korea/es/2003(H5N1))
AGW27062.1 A/wild bird/Hong Kong/07035-1/2011(H5N1))
ABB20232.1 A/duck/Hong Kong/319/1978(H2N2))
AAY87411.1 A/mallard/Postdam/178-4/1983(H2N2))
AFM09392.1 A/emperor goose/Alaska/44297-260/2007(H2N2))
AAT78585.1 A/chicken/British Columbia/GSC human B/04(H7N3))
ACO95649.1 A/avian/Delaware Bay/226/2006(H7N3))
ADP07287.1 A/American green-winged teal/California/44242-906/2007(H7N3))
AFY07081.1 A/American black duck/New Brunswick/02490/2007(H7N3))
ACA25321.1 A/American black duck/NB/2538/2007(H7N3))
ABI84697.1 A/turkey/Minnesota/1/1988(H7N9))
AHZ42540.1 A/mallard/Minnesota/AI09-3770/2009(H7N9))
AGG53455.1 A/wild duck/Korea/SH19-47/2010(H7N9))
AGQ81933.1 A/duck/Jiangxi/3096/2009(H7N9))
AGQ81979.1 A/chicken/Rizhao/713/2013(H7N9))
AHD25274.1 A/chicken/Jiangsu/1021/2013(H7N9))
20
15
10
5
0

Fig. 5. Phylogenetic tree of 35 Influenza A viruses based on neuraminidase by our method. The dataset includes 5 groups: H1N1 (blue), H5N1 (red), H2N2 (purplish red), H7N3 (green), H7N9 (black). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Prediction accuracy rates and running time for our method and ClustalW. The accuracy rates are evaluated by leave-one-out and first nearest neighbor methods. The number of subjects in each dataset is marked in a bracket in the first column. The third column represents the running time to get the numerical vectors using our method. The fifth column represents the time to complete multiple alignment based on ClustalW.

Dataset	Our method	Running time	ClustalW	Running time
Beta-globin (88)	94.3%	0.06 s	84.1%	5 s
HRV (114)	100%	0.68 s	100%	21 min
Influenza data1 (35)	100%	0.04 s	97.1%	7 s
Mammalian mitochondria (79)	97.5%	0.06 s	97.5%	22 s
Influenza data2 (1163)	100%	1.33 s	99.7%	175 min

Mammalian mitochondria dataset including 79 proteins, the two methods produce 97.5% accuracy rate.

We further validate our method by an independent dataset (Influenza data2) consisting of 1163 Influenza A viruses isolated in China. The dataset was downloaded from Global Initiative on Sharing Avian Influenza Data (GISAID) database (http://platform.gisaid.org/epi3/frontend#377f5). The dataset includes 13 species: H5N6, H5N1, H7N9, H1N1, H6N2, H3N8, H3N2, H4N6, H5N5, H10N3 and H7N3. The summary for these sequences is listed in Table 4. The first column of the table presents the names of subtypes. The second column presents the number of strains in each subtype. The third, fourth, and fifth columns are the minimum length, mean length and maximum length of NA proteins in each subtype group. The neuraminidase protein encoded by the neuraminidase (NA) gene is utilized to evaluate the predication accuracy and construct the phylogenetic tree of these viruses. We also employ the 1-NN method to compute the predication correctness rate. For an influenza A virus for example H9N2, if its nearest neighbor is a virus belonging to the same subtype, i.e. H9N2, we think the prediction is correct. As shown in Table 3, our prediction accuracy is 100% for

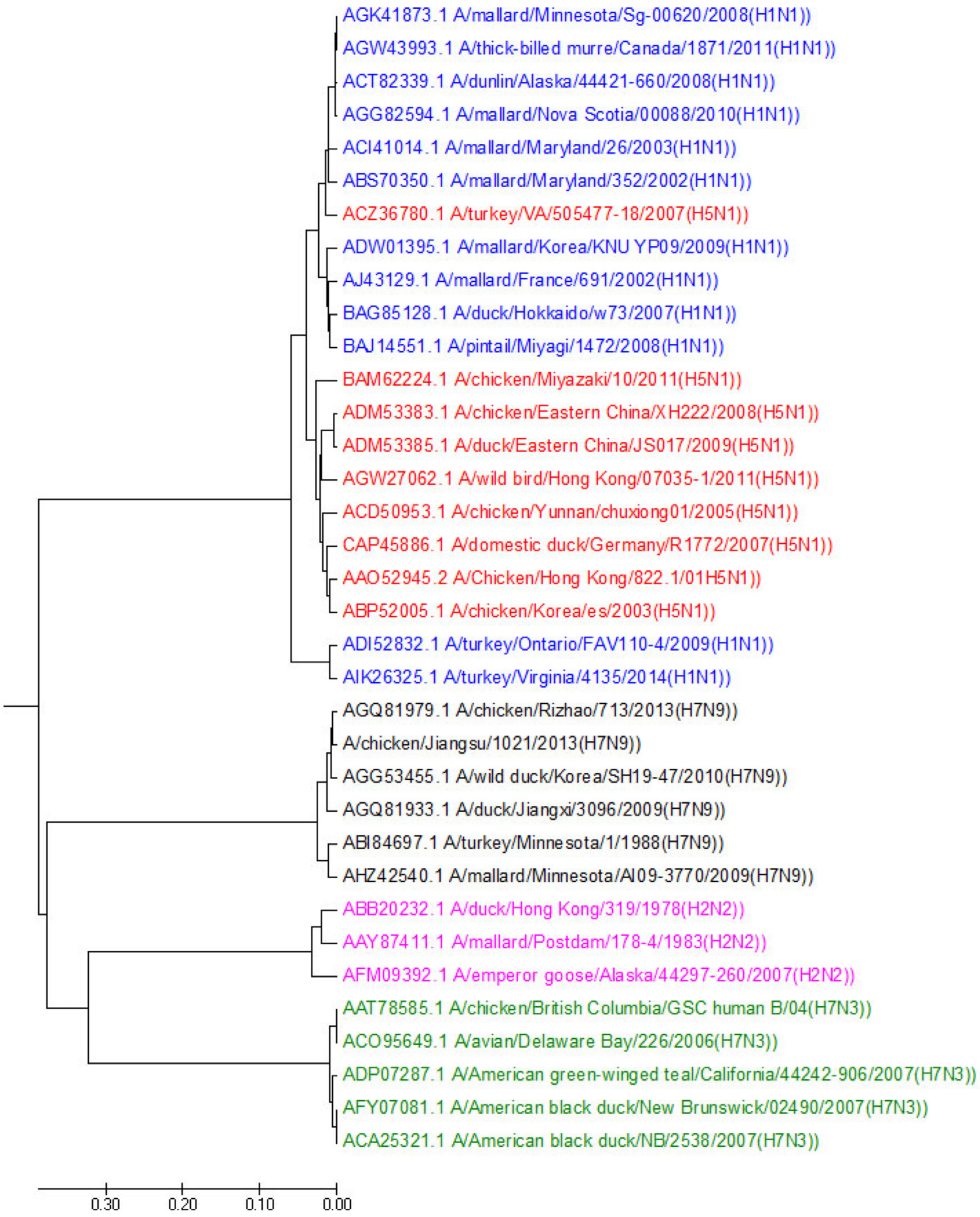
AGK41873.1 A/mallard/Minnesota/Sg-00620/2008(H1N1))
AGW43993.1 A/thick-billed murre/Canada/1871/2011(H1N1))
ACT82339.1 A/dunlin/Alaska/44421-660/2008(H1N1))
AGG82594.1 A/mallard/Nova Scotia/00088/2010(H1N1))
ACI41014.1 A/mallard/Maryland/26/2003(H1N1))
ABS70350.1 A/mallard/Maryland/352/2002(H1N1))
ACZ36780.1 A/turkey/VA/505477-18/2007(H5N1))
ADW01395.1 A/mallard/Korea/KNU YP09/2009(H1N1))
AJ43129.1 A/mallard/France/691/2002(H1N1))
BAG85128.1 A/duck/Hokkaido/w73/2007(H1N1))
BAJ14551.1 A/pintail/Miyagi/1472/2008(H1N1))
BAM62224.1 A/chicken/Miyazaki/10/2011(H5N1))
ADM53383.1 A/chicken/Eastern China/XH222/2008(H5N1))
ADM53385.1 A/duck/Eastern China/JS017/2009(H5N1))
AGW27062.1 A/wild bird/Hong Kong/07035-1/2011(H5N1))
ACD50953.1 A/chicken/Yunnan/chuxiong01/2005(H5N1))
CAP45886.1 A/domestic duck/Germany/R1772/2007(H5N1))
AAO52945.2 A/Chicken/Hong Kong/822.1/01H5N1))
ABP52005.1 A/chicken/Korea/es/2003(H5N1))
ADI52832.1 A/turkey/Ontario/FAV110-4/2009(H1N1))
AIK26325.1 A/turkey/Virginia/4135/2014(H1N1))
AGQ81979.1 A/chicken/Rizhao/713/2013(H7N9))
A/chicken/Jiangsu/1021/2013(H7N9))
AGG53455.1 A/wild duck/Korea/SH19-47/2010(H7N9))
AGQ81933.1 A/duck/Jiangxi/3096/2009(H7N9))
ABI84697.1 A/turkey/Minnesota/1/1988(H7N9))
AHZ42540.1 A/mallard/Minnesota/AI09-3770/2009(H7N9))
ABB20232.1 A/duck/Hong Kong/319/1978(H2N2))
AAY87411.1 A/mallard/Postdam/178-4/1983(H2N2))
AFM09392.1 A/emperor goose/Alaska/44297-260/2007(H2N2))
AAT78585.1 A/chicken/British Columbia/GSC human B/04(H7N3))
ACO95649.1 A/avian/Delaware Bay/226/2006(H7N3))
ADP07287.1 A/American green-winged teal/California/44242-906/2007(H7N3))
AFY07081.1 A/American black duck/New Brunswick/02490/2007(H7N3))
ACA25321.1 A/American black duck/NB/2538/2007(H7N3))
0.30
0.20
0.10
0.00

Fig. 6. Phylogenetic tree of 35 Influenza A viruses based on neuraminidase by ClustalW. The dataset includes 5 groups: H1N1 (blue), H5N1 (red), H2N2 (purplish red), H7N3 (green), H7N9 (black). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

this dataset, while the alignment approach generates 99.7% accuracy rate.

As for the computational complexity shown in Table 3, it takes about 1 s to compute our vectors for the 1163 neuraminidase sequences, while the popular multiple alignment algorithm ClustalW spends about 3 h to complete the alignment. The UPGMA tree for these viruses is also constructed by our method. As shown in Fig. 9, the 13 subtypes are clearly separated from each other. Based on ClustalW, our phylogenetic tree of the 1163 viruses are built. As shown in Fig. 10, most of the strains from a same species are clustered together. However, A/duck/Guangdong/1/1996/EPI383286 H7N3 strain is positioned with H10N3, which seems unreliable. The strains from H1N1 are divided into two major clades. Moreover, A/wildbird/Wuhan/WHHN58/2014/ EPI682990 H1N1 strain and A/duck/FuJian/JF47/2014 EPI703429 H1N1 strain are placed among H5N1 strains.

4. Discussion and conclusion

In this paper, we use three important biochemical properties of amino acids and come up with a 24 dimensional vector to compare protein sequences. Among alignment-free method for comparing proteins, our feature vector is lower in dimension. Besides, our method both runs fast and clusters proteins precisely. Compar-

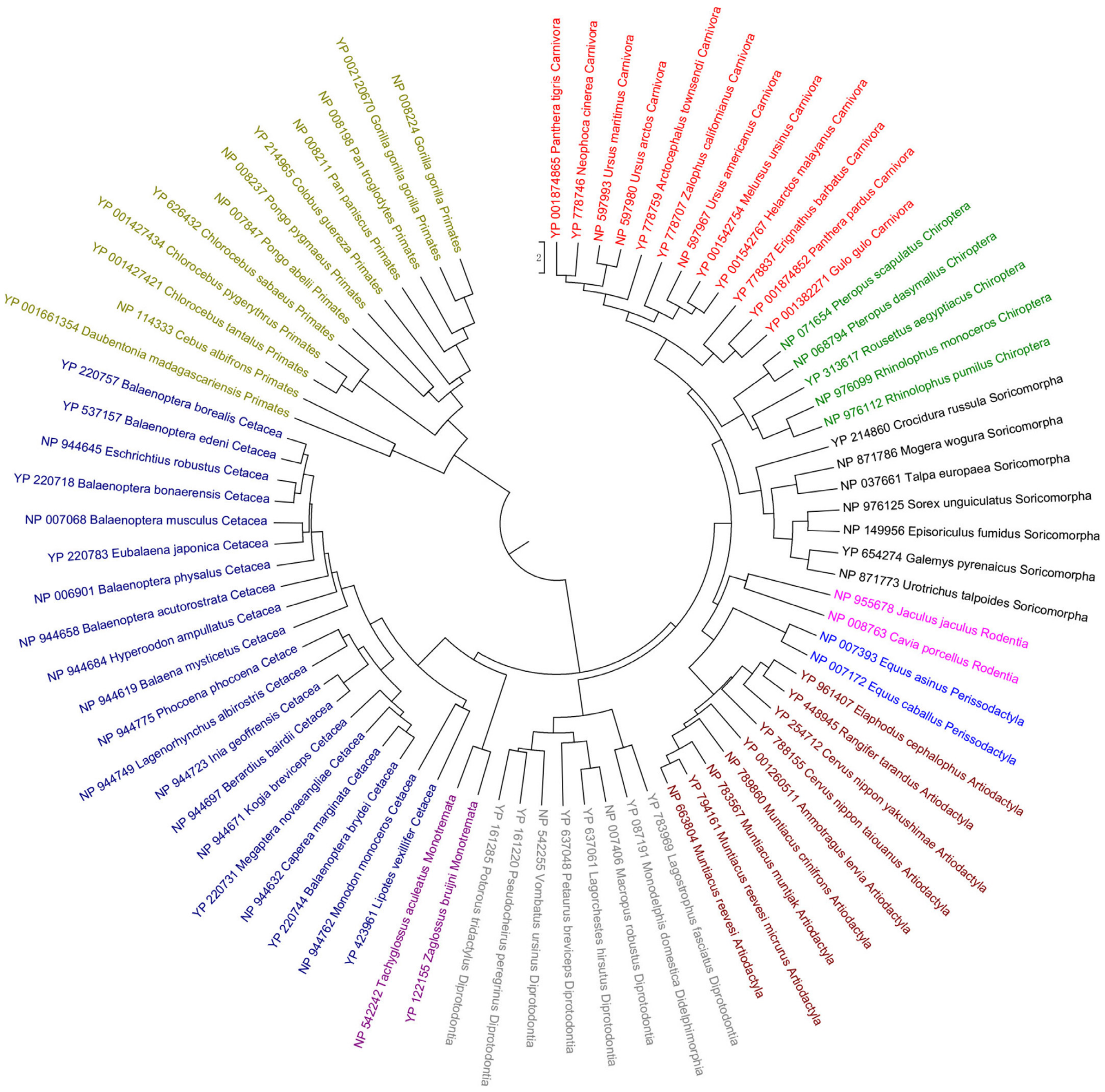
2
YP 001874865 Panthera tigris Carnivora
YP 778746 Neophoca cinerea Carnivora
NP 597993 Ursus maritimus Carnivora
NP 597980 Ursus arctos Carnivora
YP 778759 Arctocephalus townsendi Carnivora
YP 778707 Zalophus californianus Carnivora
NP 597967 Ursus americanus Carnivora
YP 001542754 Melursus ursinus Carnivora
YP 001542767 Helarctos malayanus Carnivora
YP 778837 Erignathus barbatus Carnivora
YP 001874852 Panthera pardus Carnivora
YP 001382271 Gulo gulo Carnivora
NP 071654 Pteropus scapulatus Chiroptera
NP 068794 Pteropus dasymallus Chiroptera
YP 313617 Rousettus aegyptiacus Chiroptera
NP 976099 Rhinolophus monoceros Chiroptera
NP 976112 Rhinolophus pumilus Chiroptera
YP 214860 Crocidura russula Soricomorpha
NP 871786 Mogera wogura Soricomorpha
NP 037661 Talpa europaea Soricomorpha
NP 976125 Sorex unguiculatus Soricomorpha
NP 149956 Episoriculus fumidus Soricomorpha
YP 654274 Galemys pyrenaicus Soricomorpha
NP 871773 Urotrichus talpoides Soricomorpha
NP 955678 Jaculus jaculus Rodentia
NP 008763 Cavia porcellus Rodentia
NP 007393 Equus asinus Perissodactyla
NP 007172 Equus caballus Perissodactyla
YP 961407 Elaphodus cephalophus Artiodactyla
YP 448945 Rangifer tarandus Artiodactyla
YP 254712 Cervus nippon yakushimae Artiodactyla
YP 788155 Cervus nippon taiouanus Artiodactyla
YP 001260511 Ammotragus lervia Artiodactyla
NP 789860 Muntiacus crinifrons Artiodactyla
NP 783567 Muntiacus muntjak Artiodactyla
YP 794161 Muntiacus reevesi micrurus Artiodactyla
NP 663804 Muntiacus reevesi Artiodactyla
YP 783969 Lagostrophus fasciatus Diprotodontia
YP 087191 Monodelphis domestica Didelphimorphia
NP 007406 Macropus robustus Diprotodontia
YP 637061 Lagorchestes hirsutus Diprotodontia
YP 637048 Petaurus breviceps Diprotodontia
NP 542255 Vombatus ursinus Diprotodontia
YP 161220 Pseudocheirus peregrinus Diprotodontia
YP 161285 Potorous tridactylus Diprotodontia
YP 122155 Zaglossus bruijni Monotremata
NP 542242 Tachyglossus aculeatus Monotremata
YP 423961 Lipotes vexillifer Cetacea
NP 944762 Monodon monoceros Cetacea
YP 220744 Balaenoptera brydei Cetacea
NP 944632 Caperea marginata Cetacea
YP 220731 Megaptera novaeangliae Cetacea
NP 944671 Kogia breviceps Cetacea
NP 944697 Berardius bairdii Cetacea
NP 944723 Inia geoffrensis Cetacea
NP 944749 Lagenorhynchus albirostris Cetacea
NP 944775 Phocoena phocoena Cetace
NP 944619 Balaena mysticetus Cetacea
NP 944684 Hyperoodon ampullatus Cetacea
NP 944658 Balaenoptera acutorostrata Cetacea
NP 006901 Balaenoptera physalus Cetacea
YP 220783 Eubalaena japonica Cetacea
NP 007068 Balaenoptera musculus Cetacea
YP 220718 Balaenoptera bonaerensis Cetacea
NP 944645 Eschrichtius robustus Cetacea
YP 537157 Balaenoptera edeni Cetacea
YP 220757 Balaenoptera borealis Cetacea
YP 001661354 Daubentonia madagascariensis Primates
NP 114333 Cebus albifrons Primates
YP 001427421 Chlorocebus tantalus Primates
YP 001427434 Chlorocebus pygerythrus Primates
YP 626432 Chlorocebus sabaeus Primates
NP 007847 Pongo abelii Primates
NP 008237 Pongo pygmaeus Primates
YP 214965 Colobus guereza Primates
NP 008211 Pan paniscus Primates
NP 008198 Pan troglodytes Primates
YP 002120670 Gorilla gorilla gorilla Primates
NP 008224 Gorilla gorilla Primates

Fig. 7. Phylogenetic tree of 79 mammals based on by our method. The dataset includes 10 groups: *Carnivora* (red), *Chiroptera* (green), *Soricomorpha* (black), *Rodentia* (purplish red), *Perissodactyla* (blue), *Artiodactyla* (maroon), *Diprotodontia* (gray), *Monotremata* (purple), *Cetacea* (navy blue), *Primates* (olive). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ing with ClustalW, our approach is not only able to handle large data set such as mammalian mitochondria dataset, but also capable to analyze long protein sequences as in the Human rhinovirus dataset. Although it is difficult to consider all chemical properties of amino acids, our new method based on three crucial properties performs very well on diverse datasets. In addition, vector models can be directly used to predict various attributes of proteins in many different areas. For example, the PseAAC or general PseAAC approaches were applied to many fields such as subcellular localization of proteins (Chou and Shen, 2007b), membrane protein types (Chou and Shen, 2007a), and enzyme family class (Chou, 2005). Our vector method has the potential to classify various proteins. Like the PseAAC method, our vectors can be imported into most of the current machine-learning algorithms which can only handle vectors but not sequence samples as elucidated in a recent review (Chou, 2015).

For each group of amino acids, we define three features to describe it: the count, the average of its position and the variability of its position. The count of these amino acids is a common feature to present composition of protein sequences. Moreover, the three features are able to distinguish different clusters of proteins. For example, the three features of α in the four clades HEV-c, HRV-A, HRV-B, HRV-C of HRV dataset exhibit much difference. The number of α, i.e. n_α in proteins in the four clades are respectively

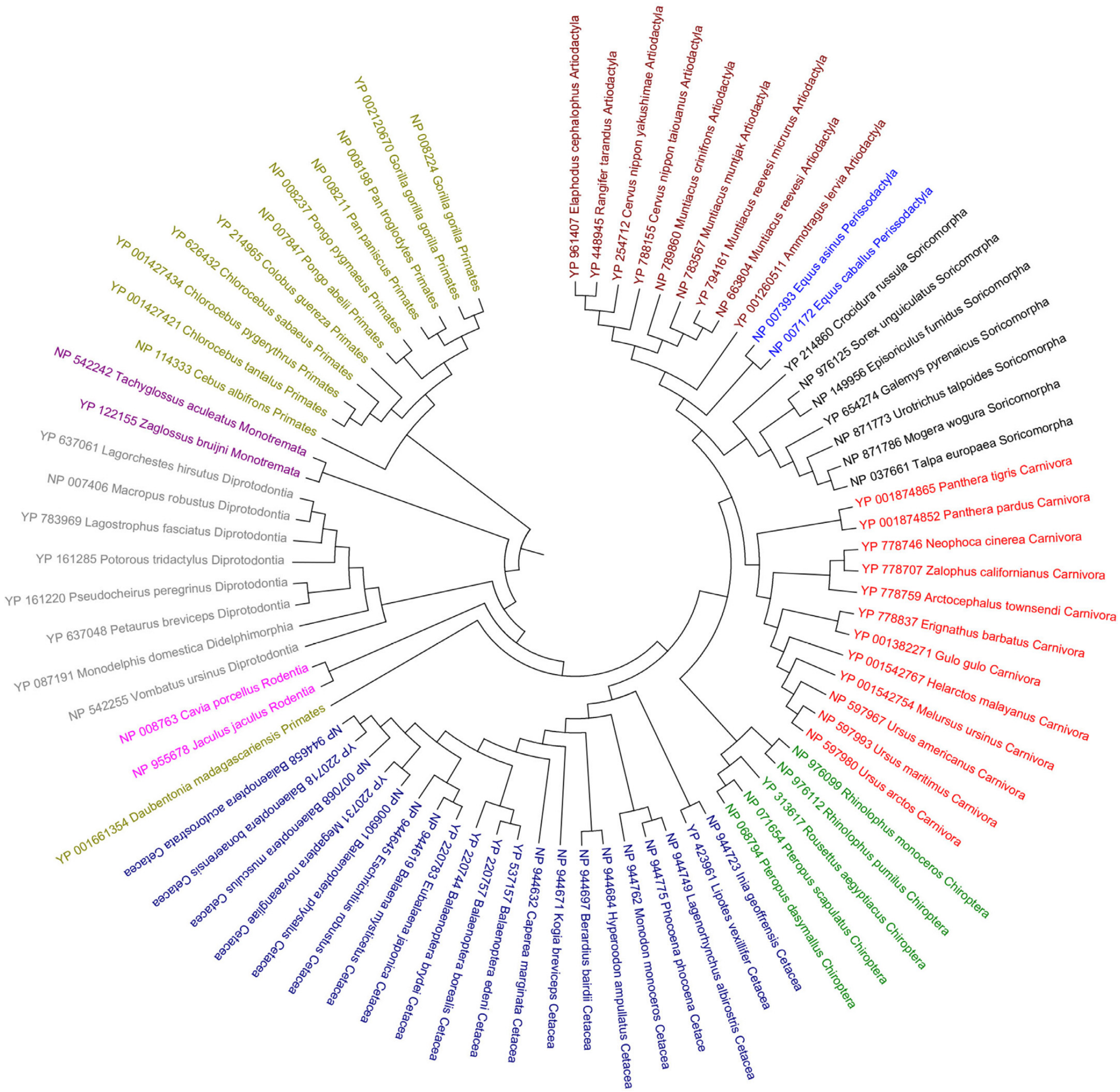
YP 961407 Elaphodus cephalophus Artiodactyla
YP 448945 Rangifer tarandus Artiodactyla
YP 254712 Cervus nippon yakushimae Artiodactyla
YP 788155 Cervus nippon taiouanus Artiodactyla
NP 789860 Muntiacus crinifrons Artiodactyla
NP 783567 Muntiacus muntjak Artiodactyla
YP 794161 Muntiacus reevesi micrurus Artiodactyla
NP 663804 Muntiacus reevesi Artiodactyla
YP 001260511 Ammotragus lervia Artiodactyla
NP 007393 Equus asinus Perissodactyla
NP 007172 Equus caballus Perissodactyla
YP 214860 Crocidura russula Soricomorpha
NP 976125 Sorex unguiculatus Soricomorpha
NP 149956 Episoriculus fumidus Soricomorpha
YP 654274 Galemys pyrenaicus Soricomorpha
NP 871773 Urotrichus talpoides Soricomorpha
NP 871786 Mogera wogura Soricomorpha
NP 037661 Talpa europaea Soricomorpha
YP 001874865 Panthera tigris Carnivora
YP 001874852 Panthera pardus Carnivora
YP 778746 Neophoca cinerea Carnivora
YP 778707 Zalophus californianus Carnivora
YP 778759 Arctocephalus townsendi Carnivora
YP 778837 Erignathus barbatus Carnivora
YP 001382271 Gulo gulo Carnivora
YP 001542767 Helarctos malayanus Carnivora
YP 001542754 Melursus ursinus Carnivora
NP 597967 Ursus americanus Carnivora
NP 597993 Ursus maritimus Carnivora
NP 597980 Ursus arctos Carnivora
NP 976099 Rhinolophus monoceros Chiroptera
NP 976112 Rhinolophus pumilus Chiroptera
YP 313617 Rousettus aegyptiacus Chiroptera
NP 071654 Pteropus scapulatus Chiroptera
NP 068794 Pteropus dasymallus Chiroptera
NP 944723 Inia geoffrensis Cetacea
YP 423961 Lipotes vexillifer Cetacea
NP 944749 Lagenorhynchus albirostris Cetacea
NP 944775 Phocoena phocoena Cetacea
NP 944762 Monodon monoceros Cetacea
NP 944684 Hyperoodon ampullatus Cetacea
NP 944697 Berardius bairdii Cetacea
NP 944671 Kogia breviceps Cetacea
NP 944632 Caperea marginata Cetacea
YP 537157 Balaenoptera edeni Cetacea
YP 220757 Balaenoptera borealis Cetacea
YP 220744 Balaenoptera brydei Cetacea
YP 220783 Eubalaena japonica Cetacea
NP 944619 Balaena mysticetus Cetacea
NP 944645 Eschrichtius robustus Cetacea
NP 006901 Balaenoptera physalus Cetacea
YP 220731 Megaptera novaeangliae Cetacea
NP 007068 Balaenoptera musculus Cetacea
YP 220718 Balaenoptera bonaerensis Cetacea
NP 944658 Balaenoptera acutorostrata Cetacea
YP 001661354 Daubentonia madagascariensis Primates
NP 955678 Jaculus jaculus Rodentia
NP 008763 Cavia porcellus Rodentia
NP 542255 Vombatus ursinus Diprotodontia
YP 087191 Monodelphis domestica Didelphimorphia
YP 637048 Petaurus breviceps Diprotodontia
YP 161220 Pseudocheirus peregrinus Diprotodontia
YP 161285 Potorous tridactylus Diprotodontia
YP 783969 Lagostrophus fasciatus Diprotodontia
NP 007406 Macropus robustus Diprotodontia
YP 637061 Lagorchestes hirsutus Diprotodontia
YP 122155 Zaglossus bruijni Monotremata
NP 542242 Tachyglossus aculeatus Monotremata
NP 114333 Cebus albifrons Primates
YP 001427421 Chlorocebus tantalus Primates
YP 001427434 Chlorocebus pygerythrus Primates
YP 626432 Chlorocebus sabaeus Primates
YP 214965 Colobus guereza Primates
NP 007847 Pongo abelii Primates
NP 008237 Pongo pygmaeus Primates
NP 008211 Pan paniscus Primates
NP 008198 Pan troglodytes Primates
YP 002120670 Gorilla gorilla gorilla Primates
NP 008224 Gorilla gorilla Primates

Fig. 8. Phylogenetic tree of 79 mammals based on ClustalW. The dataset includes 10 groups: *Carnivora* (red), *Chiroptera* (green), *Soricomorpha* (black), *Rodentia* (purplish red), *Perissodactyla* (blue), *Artiodactyla* (maroon), *Diprotodontia* (gray), *Monotremata* (purple), *Cetacea* (navy blue), *Primates* (olive). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

about 810, 770, 780 and 760 on average. The mean position of α, i.e. μ_α in proteins in the four clades are respectively about 1120, 1100, 1110 and 1085 on average. The D_2^α values in proteins in the four clades are respectively about 179, 175, 177 and 176 on average.

Our method is different from the classical k-mer method. The k-mer method computes the probabilities of occurrence of strings with k letters. However, our method contains the number of each letter. In addition, our new vectors also include the average position, variability of position of each letter and have much lower dimensions than k-mer method. We also employ the k-mer method to build the phylogenetic trees of 88 Beta-globin proteins. The Euclidian distance is utilized to measure the similarity between k-mer vectors. For the tree based on 2-mer method, the three proteins from Anura are not clustered together (Figure S1 in supporting files). For the tree with 3-mer method, besides the three proteins from Anura, the five proteins from Anseriformes are not positioned together (Figure S2 in supporting files).

Once the pairwise distances are obtained, distance based algorithms such as neighbor-joining or Fitch-Margoliash also can be used to build phylogenetic trees. As comparison, we construct the neighbor-joining tree of 35 Influenza A viruses. In this tree shown in Figure S3, all strains from the same species are placed together, which is consistent with our UPGMA tree.

Table 4
Summary of the neuraminidase (NA) proteins of 1163 influenza A viruses.

Subtype	Number	Min length	Mean length	Max length (aa)
H5N6	12	451	457.4167	460
H5N1	321	427	446.9907	460
H7N9	53	456	459.9623	465
H1N1	171	467	469.0234	470
H9N2	379	465	466.4512	469
H6N2	82	469	469	469
H3N8	10	469	469.9	470
H3N2	83	466	468.9639	469
H4N6	19	470	470	470
H6N6	6	470	470	470
H5N5	5	472	472	472
H10N3	7	469	469	469
H7N3	15	469	469	469

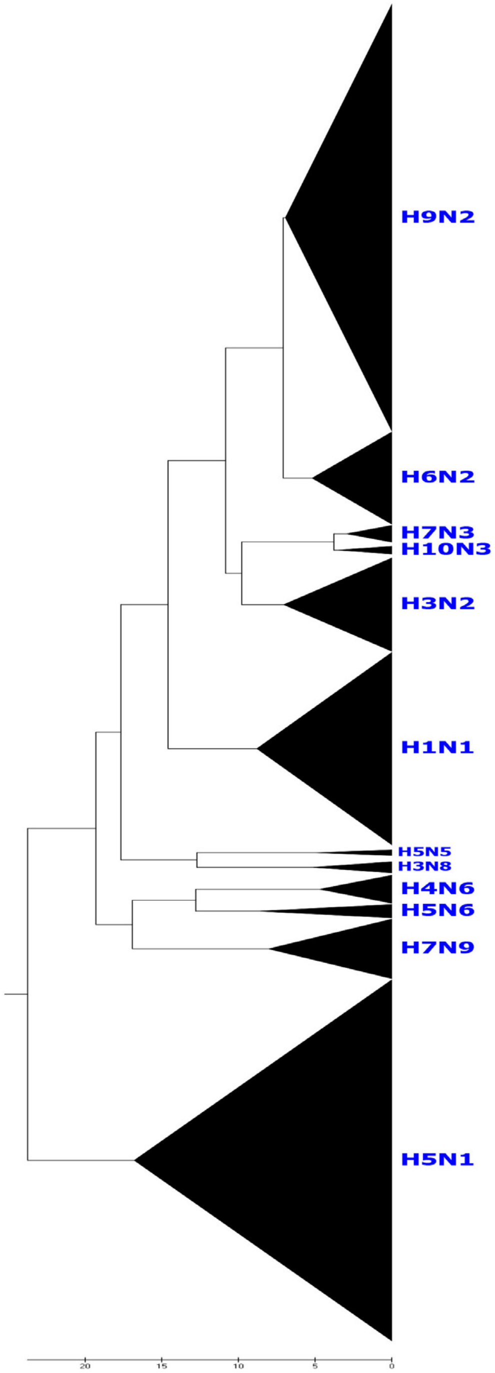

Fig. 9. Phylogenetic tree of 1163 influenza A viruses based on our method.

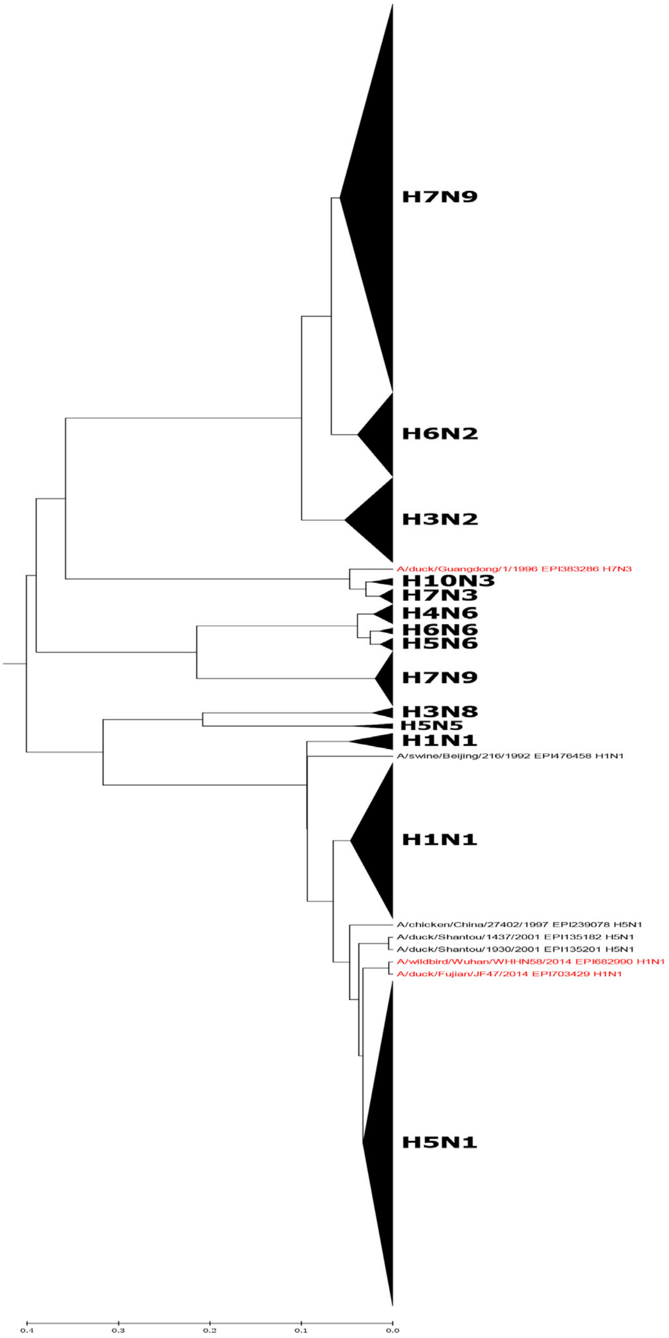

Fig. 10. Phylogenetic tree of 1163 influenza A viruses based on ClustalW.

The main limitation of our approach is the choice of certain cutoff classifying amino acids based on their values in biochemical properties. For each of the three chemical properties, we chose different cutoff according to the values of this property for the 20 amino acids. These choices perhaps can be affected by divergence of protein sequences. Further study may be implemented to explore the influence of more chemical properties and more varied cutoff. The R source code in this paper is freely available to the public upon request. User-friendly and open web-servers can provide more practical help for biologists as emphasized in two reviews (Chou, 2015; Chou and Shen, 2009). For example, some web-servers based on PseACC are useful to identify attributes of sequences (Chen et al., 2016a; Cheng et al., 2016) and special sites of sequences such as carbonylation sites (Jia et al., 2016a), RNA pseudouridine sites (Chen et al., 2016b), ysine succinylation sites

(Jia et al., 2016c), phosphorylation sites (Qiu et al., 2016b), original location of replication (Zhang et al., 2016), hydroxyproline and hydroxylysine in proteins (Qiu et al., 2016a) and the adenosine to inosine editing sites (Chen et al., 2017). We will make efforts to provide a web-server for our method in future work.

Conflict of interest statement

The authors declare no competing financial interests.

Author contributions

Stephen S.-T. Yau initiated and designed the study. Lily He and Yongkun Li developed the method. Lily He, Yongkun Li and Rong Lucy He analyzed the results. All authors wrote the manuscript and approved the final version.

Acknowledgements

This study is supported by the National Natural Science Foundation of China (31271408 to S.S.-T. Yau), Tsinghua University start up fund (to S. S.-T. Yau). Prof. Stephen Yau is grateful to National Center for Theoretical Sciences (NCTS) for providing excellent research environment while part of this research was done. The funders did not take part in study design; in collection and analysis of data; in the writing of the manuscript; in the decision to publish this manuscript.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.jtbi.2017.06.002.

References

Blaisdell, B.E., 1989. Average values of a dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a computer-generated model system. J. Mol. Evol. 29, 538–547.

Cao, D.S., Xu, Q.S., Liang, Y.Z., 2013. Propy: a tool to generate various modes of chou's pseaac. Bioinformatics 29, 960–962.

Chen, W., Ding, H., Feng, P., Lin, H., Chou, K.C., 2016a. Iacp: a sequence-based tool for identifying anticancer peptides. Oncotarget 7, 16895–16909. doi:10.18632/oncotarget.7815.

Chen, W., Feng, P., Yang, H., Ding, H., Lin, H., Chou, K.C., 2017. Irna-ai: identifying the adenosine to inosine editing sites in rna sequences. Oncotarget 8, 4208–4217.

Chen, W., Lei, T.Y., Jin, D.C., Lin, H., Chou, K.C., 2014. Pseknc: a flexible web server for generating pseudo k-tuple nucleotide composition. Anal. Biochem. 456, 53–60.

Chen, W., Lin, H., Feng, P.M., Ding, C., Zuo, Y.C., Chou, K.C., 2012. Inuc-physchem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. PLoS One 7, e47843. doi:10.1371/journal.pone.0047843.

Chen, W., Tang, H., Ye, J., Lin, H., Chou, K.C., 2016b. Irna-pseu: identifying rna pseudouridine sites. Molecul. Ther. Nucleic Acids.

Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., Chou, K.C., 2015. Pseknc-general: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics 31, 119–120.

Cheng, X., Zhao, S.G., Xiao, X., Chou, K.C., 2016. Iatc-misf: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. Bioinformatics doi:10.1093/bioinformatics/btw644.

Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins Struct. Funct. Bioinform. 43 (3), 246–255.

Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19. doi:10.1093/bioinformatics/bth466.

Chou, K.C., 2009. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr. Proteomics 6 (4), 262–274. (13)

Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273, 236–247. doi:10.1016/j.jtbi.2010.12.024.

Chou, K.C., 2015. Impacts of bioinformatics to medicinal chemistry. Medicinal chemistry (Shariqah (United Arab Emirates)) 11, 218–234.

Chou, K.C., Shen, H.B., 2007. Memtype-2l: a web server for predicting membrane proteins and their types by incorporating evolution information through psepssm. Biochem. Biophys. Res. Commun. 360, 339–345. doi:10.1016/j.bbrc.2007.06.027.

Chou, K.C., Shen, H.B., 2007. Review: recent progresses in protein subcellular location prediction. anal. biochem. 370, 1–16. Anal. Biochem. 370 (1), 1–16.

Chou, K.C., Shen, H.B., 2009. Review : recent advances in developing web-servers for predicting protein attributes. Nat. Sci. 1 (2), 63–92.

Chou, K.C., Zhang, C.T., 1995. Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349. doi:10.3109/10409239509083488.

Deng, M., Yu, C., Liang, Q., He, R.L., Yau, S.S.T., 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. PLoS One 6 (3), e17293.

Du, P., Gu, S., Jiao, Y., 2014. Pseaac-general: fast building various modes of general form of chou's pseudo-amino acid composition for large-scale protein datasets. Int. J. Mol. Sci. 15, 3495–3506. doi:10.3390/ijms15033495.

Du, P., Wang, X., Xu, C., Gao, Y., 2012. Pseaac-builder: a cross-platform stand-alone program for generating various special chou's pseudo-amino acid compositions. Anal. Biochem. 425, 117–119. doi:10.1016/j.ab.2012.03.015.

Edgar, R.C., 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. doi:10.1093/nar/gkh340.

Gan, H.H., Perlow, R.A., Roy, S., Ko, J., Wu, M., Huang, J., Yan, S., Nicoletta, A., Vafai, J., Sun, D., Wang, L., Noah, J.E., Pasquali, S., Schlick, T., 2002. Analysis of protein sequence/structure similarity relationships. Biophys. J. 83, 2781–2791.

Grantham, R., 1974. Amino acid difference formula to help explain protein evolution. Science 185, 862–864.

Hoang, T., Yin, C., Zheng, H., Yu, C., Lucy, H.R., Yau, S.S., 2015. A new method to cluster dna sequences using fourier power spectrum. J. Theor. Biol. 372, 135–145.

Huang, H.H., Yu, C., Zheng, H., Hernandez, T., Yau, S.C., He, R.L., Yang, J., Yau, S.S.T., 2014. Global comparison of multiple-segmented viruses in 12-dimensional genome space. Mol. Phylogenet. Evol. 81, 29–36.

Jacobs, S.E., Lamson, D.M., St George, K., Walsh, T.J., 2013. Human rhinoviruses. Clin. Microbiol. Rev. 26, 135–162. doi:10.1128/CMR.00077-12.

Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.C., 2015. Ippi-esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into pseaac. J. Theor. Biol. 377, 47–56. doi:10.1016/j.jtbi.2015.04.011.

Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.C., 2016. Icar-psecp: identify carbonylation sites in proteins by monte carlo sampling and incorporating sequence coupled effects into general pseaac. Oncotarget 7, 34558–34570. doi:10.18632/oncotarget.9148.

Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.C., 2016. Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition. J. Biomol. Struct. Dyn. 34, 1946–1961. doi:10.1080/07391102.2015.1095116.

Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.C., 2016. Psuc-lys: predict lysine succinylation sites in proteins with pseaac and ensemble random forest approach. J. Theor. Biol. 394, 223–230. doi:10.1016/j.jtbi.2016.01.020.

Jiao, Y.S., Du, P.F., 2017. Predicting protein submitochondrial locations by incorporating the positional-specific physicochemical properties into chou's general pseudo-amino acid compositions. J. Theor. Biol. 416, 81–87. doi:10.1016/j.jtbi.2016.12.026.

de Jong, M.D., Bach, V.C., Phan, T.Q., Vo, M.H., Tran, T.T., Nguyen, B.H., Beld, M., Le, T.P., Truong, H.K., Nguyen, V.V.C., Tran, T.H., Do, Q.H., Farrar, J., 2005. Fatal avian influenza a (h5n1) in a child presenting with diarrhea followed by coma. N. Engl. J. Med. 352, 686–691. doi:10.1056/NEJMoa044307.

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. doi:10.1073/pnas.0913033107.

Katoh, K., Misawa, K., Kuma, K.i., Miyata, T., 2002. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Res. 30, 3059–3066.

Khan, M., Hayat, M., Khan, S.A., Iqbal, N., 2017. Unb-dpc: identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into chou's general pseaac. J. Theor. Biol. 415, 13–19. doi:10.1016/j.jtbi.2016.12.004.

Kumar, R., Srivastava, A., Kumari, B., Kumar, M., 2015. Prediction of lactamase and its class by chou's pseudo-amino acid composition and support vector machine. J. Theor. Biol. 365, 96–103. doi:10.1016/j.jtbi.2014.10.008.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.

Li, Y., He, L., He, R.L., Yau, S.S.T., 2017. Zika and flaviviruses phylogeny based on the alignment-free natural vector method. DNA Cell Biol. 36, 109–116. doi:10.1089/dna.2016.3532.

Li, Y., Tian, K., Yin, C., He, R.L., Yau, S.S.T., 2016. Virus classification in 60-dimensional protein space. Mol. Phylogenet. Evol. 99, 53–62.

Lin, S.X., Lapointe, J., 2013. Theoretical and experimental biology in onea symposium in honour of professor kuo-chen chous 50th anniversary and professor richard giegs 40th anniversary of their scientific careers. J. Biomed. Sci. Eng. 06 (4), 435–442.

Liu, B., Liu, F., Fang, L., Wang, X., Chou, K.C., 2015. Repdna: a python package to generate various modes of feature vectors for dna sequences by incorporating user-defined physicochemical properties and sequence-order effects. Bioinformatics 31, 1307–1309. doi:10.1093/bioinformatics/btu820.

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., Chou, K.C., 2015. Pse-in-one: a web server for generating various modes of pseudo components of dna, rna, and protein sequences. Nucleic Acids Res. 43, W65–W71. doi:10.1093/nar/gkv458.

Meher, P.K., Sahu, T.K., Saini, V., Rao, A.R., 2017. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into chou's general pseaac. Sci. Rep. 7, 42362. doi:10.1038/srep42362.

Mondal, S., Pai, P.P., 2014. Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. J. Theor. Biol. 356, 30–35. doi:10.1016/j.jtbi.2014.04.006.

Palmenberg, A.C., Spiro, D., Kuzmickas, R., Wang, S., Djikeng, A., Rathe, J.A., Fraserliggett, C.M., Liggett, S.B., 2009. Sequencing and analyses of all known human rhinovirus genomes reveal structure and evolution. Science 324 (5923), 55–59.

Pearl, F., Todd, A.E., Bray, J.E., Martin, A.C., Salamov, A.A., Suwa, M., Swindells, M.B., Thornton, J.M., Orengo, C.A., 2000. Using the cath domain database to assign structures and functions to the genome sequences. Biochem. Soc. Trans. 28, 269–275.

Qi, J., Luo, H., Hao, B., 2004. Cvtree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res. 32, W45–W47. doi:10.1093/nar/gkh362.

Qiu, W.R., Sun, B.Q., Xiao, X., Xu, Z.C., Chou, K.C., 2016. Ihyd-psecp: identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general pseaac. Oncotarget 7, 44310–44321. doi:10.18632/oncotarget.10027.

Qiu, W.R., Xiao, X., Xu, Z.C., Chou, K.C., 2016. Iphos-pseen: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. Oncotarget 7, 51270–51283. doi:10.18632/oncotarget.9987.

Rackovsky, S., 2009. Sequence physical properties encode the global organization of protein structure space. Proc. Natl. Acad. Sci. U.S.A. 106, 14345–14348. doi:10.1073/pnas.0903433106.

Salichos, L., Rokas, A., 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497, 327–331. doi:10.1038/nature12130.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. Mega6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30 (12), 2725–2729.

Tobe, S.S., Kitchener, A.C., Linacre, A.M.T., 2010. Reconstructing mammalian phylogenies: a detailed comparison of the cytochrome b and cytochrome oxidase subunit i mitochondrial genes. PloS One 5, e14156. doi:10.1371/journal.pone.0014156.

Tong, S., Zhu, X., Li, Y., Shi, M., Zhang, J., Bourgeois, M., Yang, H., Chen, X., Recuenco, S., Gomez, J., Chen, L.M., Johnson, A., Tao, Y., Dreyfus, C., Yu, W., McBride, R., Carney, P.J., Gilbert, A.T., Chang, J., Guo, Z., Davis, C.T., Paulson, J.C., Stevens, J., Rupprecht, C.E., Holmes, E.C., Wilson, I.A., Donis, R.O., 2013. New world bats harbor diverse influenza a viruses. PLoS Pathog. 9, e1003657. doi:10.1371/journal.ppat.1003657.

Ulitsky, I., Burstein, D., Tuller, T., Chor, B., 2006. The average common substring approach to phylogenomic reconstruction. J. Comput. Biol. 13, 336–350. doi:10.1089/cmb.2006.13.336.

Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison-a review. Bioinformatics 19 (4), 513–523.

Webster, R.G., Bean, W.J., Gorman, O.T., Chambers, T.M., Kawaoka, Y., 1992. Evolution and ecology of influenza a viruses. Microbiol. Rev. 56, 152–179.

Wimley, W.C., White, S.H., 1996. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat. Struct. Biol. 3, 842–848.

Woese, C.R., Dugre, D.H., Dugre, S.A., Kondo, M., Saxinger, W.C., 1966. On the fundamental nature and evolution of the genetic code. Cold Spring Harb. Symp. Quant. Biol. 31, 723–736.

Wu, Z.C., Xiao, X., Chou, K.C., 2010. 2D-mh: a web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. J. Theor. Biol. 267, 29–34. doi:10.1016/j.jtbi.2010.08.007.

Xie, X.H., Yu, Z.G., Han, G.S., Yang, W.F., Anh, V., 2015. Whole-proteome based phylogenetic tree construction with inter-amino-acid distances and the conditional geometric distribution profiles. Mol. Phylogenet. Evol. 89, 37–45. doi:10.1016/j.ympev.2015.04.008.

Yau, S.S.T., Yu, C., He, R., 2008. A protein map and its application. DNA Cell Biol. 27, 241–250. doi:10.1089/dna.2007.0676.

Yin, C., Chen, Y., Yau, S.T., 2014. A measure of dna sequence similarity by fourier transform with applications on hierarchical clustering. J. Theor. Biol. 359 (24), 18–28.

Yu, C., He, R.L., Yau, S.S.T., 2013. Protein sequence comparison based on k-string dictionary. Gene 529, 250–256. doi:10.1016/j.gene.2013.07.092.

Zhang, C.J., Tang, H., Li, W.C., Lin, H., Chen, W., Chou, K.C., 2016. Iori-human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. Oncotarget 7, 69783–69793. doi:10.18632/oncotarget.11975.

Zhong, W.Z., Zhou, S.F., 2014. Molecular science for drug development and biomedicine. Int. J. Mol. Sci. 15, 20072–20078. doi:10.3390/ijms151120072.