MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **90** (2023) 357–380
ISSN: 0340–6253
doi: 10.46793/match.90-2.357Q

Protein Sequence Comparison Method Based on 3-ary Huffman Coding

Zhaohui Qi[a,b,*], **Yingqiang Ning**[a,b], **Yinmei Huang**[b]

[a] *Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing*

[b] *College of Information Science and Engineering Hunan Normal University, Changsha 410081, People's Republic of China*

zhqi_wy2013@163.com, ningyingqiang2021@qq.com, echo_hym99@163.com

(Received October 21, 2022)

Abstract

Based on 3-ary Huffman coding algorithm, we propose a digital mapping method of protein sequence. Firstly, a 3-ary Huffman tree is defined by the frequency characteristic of 20 amino acids in given protein sequences. The 0-2 codes of 20 amino acids constructed by the 3-ary Huffman tree can convert long protein sequences into one-to-one 0-2 digital sequences. According to the frequency characteristic and the distribution information of 0-2 codes of 20 amino acids in the 0-2 digital sequences, we design the 40-dimensional vectors to characterize the protein sequences. Next, the proposed digital mapping method is used to perform three separate applications, similarity comparison of nine ND6 proteins, evolutionary trend analysis of the 2009 pandemic Human influenza A (H1N1) viruses from January 2020 to June 2022, and the evolution analysis of 95 coronavirus genes. The results illustrate the utility of the proposed method.

1 Introduction

Biological sequence comparison is the most fundamental part of computational biology and bioinformatics. Traditional alignment methods by

*Corresponding author.

using dynamic programming and regression algorithms, are faced with insurmountable barriers in processing the large-scale sequence data [1] and selecting a satisfactory scoring scheme of pair-wise unit alignments [2]. Facing the explosive growth of biological sequence, such as DNA or protein sequences, many alignment-free methods have been proposed over the past decades by considering the graphical or numerical characteristics based on static mapping, statistics, geometry, physicochemical attributes of amino acids, and so on.

As for an important kind of alignment-free method, graphical representation of DNA sequences or Protein sequences provides intuitive description of sequences by visualization approaches. The early graphical representation of biological sequences is the visualization of DNA sequences, such as Hamori in 1983 [3], Hamori in 1989 [4] and Jeffrey in 1990 [5]. Afterwards, researchers developed more graphical representation methods of DNA sequences, like Nandy [6], Bielińska-Waż [7, 8], Liao [9, 10], Randić [11, 12], Jaklic [13], Qi [14], and so on. These methods effectively provide direct insights into local and global characteristics of DNA sequences. Comparing with DNA sequence, graphical representation of protein sequence is much more difficult because of twenty amino acids. The early graphical methods of protein sequence were proposed by Randić et al. [15, 16] in 2004. After that, more graphical representation methods of protein sequences were developed, such as Randić [17], Bai and Wang [18], He [19], and so on. Furthermore, researchers put forward some graphical representation methods for protein sequence by considering physicochemical properties of amino acids [20–23].

Another alternative kind of alignment-free method is the digital mapping methods of DNA sequences or Protein sequences without using graphical representation. Some digital methods numerically convert biological sequences into some characteristic vectors or matrices closely related to biology evolution. The evolution distance among sequences can be computed by some distance measures of vectors or matrices, such as the Euclidean distance [24], the Cosine distance [25], the Mahalanobis distance [26], and so on. An effective construction of the characteristic vectors or matrices is based on the statistical properties of string distribution [27, 28], which is

a quantitative factor of numerical representation of sequences. Especially, k-string-distribution methods have received extensive attention from researchers, such as k-word interval model [29], word frequencies [30], shortest absent word [31], etc. Other some alignment-free methods are proposed by researchers from new perspectives, like the numbers of adjacent Amino Acids [32], 2-step Markov Model [33], Markov Chain and Information Entropy [34]. These alignment-free methods without graphical representation provide more effective choices for evolutionary analysis and sequence comparison of biological sequences.

In this paper we proposed a new alignment-free method for protein sequence based on 3-ary Huffman coding algorithm. It has two main advantages. One is that there is no circuit or degeneracy, so that the correspondence between proteins sequences and digital sequences is one to one. The other is that the method is rapid while it assures the validity because the whole process does not relate to complex alignment algorithm. By the algorithm the 0-2 codes of 20 amino acids can convert a long protein sequence into a one-to-one 0-2 digital sequence. The unique 0-2 digital sequence of a protein sequence provides a basis for graphical representation of the protein sequence and sequence feature extraction. Next multicomponent vectors derived from 0-2 digital sequence are in making similarity/dissimilarity comparison of protein sequences. The evolutionary analysis results of three independent data sets show the efficiency of our method.

2 Methods

2.1 *k*-ary Huffman coding

Huffman coding was developed by David A. Huffman, and published in the 1952 paper "A Method for the Construction of Minimum-Redundancy Codes" [35]. It uses a specific method for choosing the representation for each symbol, resulting in a prefix code. Prefix codes express the most common characters using shorter strings of bits than are used for less common source symbols. In [35] professor Huffman further generalized his

method, which was named as k-ary Huffman coding.

The k-ary Huffman algorithm uses the alphabet set $\{0, 1, \cdots, k-1\}$ to encode message and build a k-ary tree. If k equals 2, the algorithm applies as for binary codes. Note that for k greater than 2, not all sets of source words can properly form a k-ary tree for Huffman coding. In this case, it needs to add additional 0-probability placeholders. This is because the k-ary tree must form an n to 1 contractor. If n is the number of source symbols, then $(n-1)/(k-1)$ must be an integer. It needs to add additional $k-1-(n-1)/(k-1)$ 0-probability placeholders. For binary coding, any $n-sized$ set can form such a 2 to 1 contractor because $(n-1)/(2-1)$ is an integer.

To code n source symbols to k-ary Huffman codes, it needs to create a k-ary tree of nodes. A node of k-ary tree of nodes can be either a leaf node or an internal node. Initially, the n symbols form a set Γ of n nodes. All nodes are leaf nodes, which contain the symbol itself, the weight (frequency of appearance) of the symbol. A k-ary tree is generated from left to right taking the k least probable symbols in the set Γ and putting them together to form another equivalent symbol having a probability which is equal to the sum of the k symbols. Next the k symbols are removed from the set Γ. The equivalent symbol is inserted the set Γ and becomes its new node. The set Γ has $n-k+1$ nodes. The process is repeated until there is just one equivalent node in set Γ. A k-ary tree of nodes is created successfully. As a common convention, except for all the leaf nodes, for each internal node the numbers $0, 1, 2, \cdots, k-1$ are assigned respectively to its children from left to right child nodes. k-ary Huffman code of a source symbol is the numbers $0, 1, 2, \cdots, k-1$ sequences starting from the root node to the source symbol itself.

2.2 3-ary Huffman codes for 20 amino acids

A protein sequence is usually represented as a linear sequence composed of symbolics from the set $\Phi = \{G, A, T, S, P, V, L, I, M, F, Y, W, D, E, N, Q, K, R, H, C\}$. The symbol $G, A, T, S, P, V, L, I, M, F, Y, W, D, E, N, Q, K, R, H$ and C denote amino acid Glycine, Alanine, Threonine, Serine, Proline, Valine, Leucine, Isoleucine, Methionine, Phenylalanine, Tyrosine,

Tryptophan, Aspartic, Glutamic, Aspargine, Glutamine, Lysine, Arginine, Histidine and Cysteine, respectively. It is particularly difficult to get overall characteristics of a long protein sequence. It is a useful method in making a mapping between a long protein sequence and a graphical representation. Here we introduce the k-ary Huffman tree to code the 20 amino acids.

A long protein sequence is considered as a symbol source. The source has 20 different symbols $\{G, A, T, S, P, V, L, I, M, F, Y, W, D, E, N, Q, K, R, H, C\}$. Their frequencies are $G_f, A_f, T_f, S_f, P_f, V_f, L_f, I_f, M_f, F_f, Y_f, W_f, D_f, E_f, N_f, Q_f, K_f, R_f, H_f$ and C_f, respectively. Based on the k-ary Huffman coding method, a k-ary Huffman tree for the protein sequence is created. The tree can be read backward, from left to right, assigning different numbers (number '0', the far left child with a least probability; number '1', the next child; ...; number '$k-1$', the far right child with a max probability) to different branches. Then we can get the k-ary Huffman codes.

As for 20 amino acids, in theory the value of parameter k maybe have 19 different values, $2, 3, \cdots$, and 20. The larger the value of k is, the more digital symbols the k-ary Huffman codes has. This is not conducive to the 2D graphical representation. However, the smaller the value of k is, the longer the code length becomes. Coding efficiency will become lower. Therefore, considering the graphical representation and efficiency, here we consider the 3-ary Huffman coding.

We give an example to illustrate the 3-ary Huffman coding for coding 20 amino acids. This is the protein sequence of ND5 gene for Human (H. sapiens, YP_003024036.1). For simplification, only first 100 amino acids $(1 \ldots 100)$ is listed,

MTMYATMTTLALTSLIPPILGALINPNKKNSYPHYVKSIIASTFII
SLFPTTMFMCLDQETIISNWHWATTQTTQLSLSFKLDYFSMTFIPV
ALFVTWSI.

Table 1. Statistical frequencies of ND5 gene for Human (H. sapiens, YP_003024036.1)

G	A	T	S	P	V	L	I	M	F
26	44	65	49	32	15	15	104	26	38
Y	W	D	E	N	Q	K	R	H	C
16	12	11	9	33	20	21	8	14	6

Table 2. The 3-ary Huffman codes of 20 amino acids within ND5 gene for Human

G	A	T	S	P	V	L	I	M	F
200	01	12	02	210	002	22	10	201	212
Y	W	D	E	N	Q	K	R	H	C
110	2022	2021	2020	211	111	112	0002	001	0001

Firstly, we compute the number of amino acids of the ND5 gene for Human, as shown in Table 1. Next, we sort the symbol frequencies and create leaf nodes. Based on the k-ary Huffman coding method, we can get a 3-ary Huffman tree for the protein sequence as shown in Fig. 1. Then the 0-2 codes of 20 amino acids within the sequence are shown in Table 2. By the 0-2 codes of Table 2 we can obtain the 0-2 sequence corresponding to the protein sequence as follows. For simplification, only the 0-2 sequence of the first 100 amino acids $(1 \ldots 100)$ is listed,

20112201110011220112122201221202221021021010222000122102112102111121122110211021000111000211202101001021221210100222212210121220121220100012220211112020121010022112022001202201121211112121112202220221211222202111021202201122121021000201222120021220220210

By the 3-ary Huffman coding method one can easily create a corresponding 0-2 codes of 20 amino acids for a given protein.

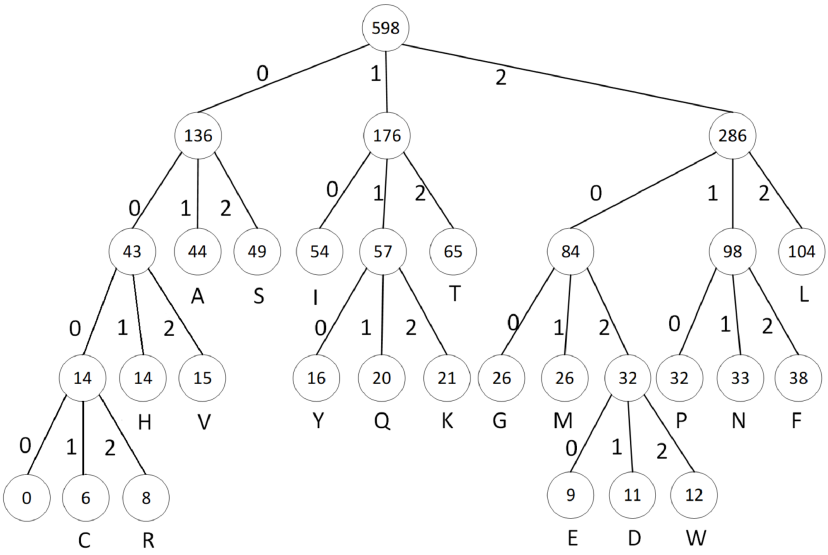

Figure 1. The 3-ary Huffman tree for 20 amino acids within ND5 gene for Human

2.3 The 2D graphic representation of protein sequence based on the 3-ary Huffman coding method

According to the 3-ary Huffman codes, a protein sequence is mapped into a sequence consisting of bit "0", "1" and "2". As shown in Fig.2, we construct a vector-graph on x-axis and two quadrants of the Cartesian coordinate system. The vectors representing number "0", "1" and "2" are the following,

$$(1,1) \to 0,\ (0,1) \to 1,\ (1,-1) \to 2.$$

The 2D graphical representation of a protein sequence based on the 3-ary Huffman codes can be obtained by connecting all the vectors one by one. In Fig.3, we illustrate the 2D graphical curves of the ND5 gene for Human (H. sapiens, YP_003024036.1) by connecting adjacent points using the vectors sequentially as shown in Fig. 2.

Fig.3 builds up a one-to-one mapping between the ND5 gene for Human

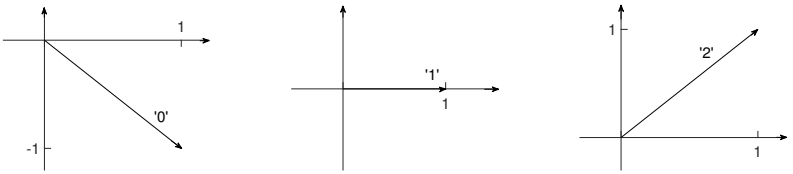

Figure 2. Graphical representations of 0, 1 and 2

and the 2D graphical curve. On one hand, for the gene sequence there is a unique 2D graphical curve corresponding to it. This is because the frequencies of amino acids in the gene sequence are fixed. Based on the fixed frequencies we can construct the corresponding 3-ary Huffman codes of the sequence. Of course, as shown in Fig. 1 one will get different 3-ary Huffman codes if he assigns 0, 1, or 2 to the branches by different way. However, the mapping between 20 amino acids and 3-ary Huffman codes will be uniquely determined once one defines a 3-ary Huffman tree. Next, by the vector-graph of Fig. 2 one can translate the sequence of numbers to a unique 2D graphical curve. Therefore, there is a unique 2D graphical curve corresponding to a given protein sequence.

To demonstrate the effectiveness of the proposed method, we give another example. There are two protein sequences from yeast Saccharomyces cerevisiae [17].

Protein I:

WTFESRNDPAKDPVILWLNGGPGCSSLTGL

Protein II:

WFFESRNDPANDPIILWLNGGPGCSSFTGL

Fig.4 gives the graphical curves of *Protein I* and *Protein II*. From an overall perspective, the two curves are very similar to each other. The graphical curves provide a very intuitive impression. Furthermore, we can easily see that there are several differences. This indicates the effectiveness of the proposed 3-ary Huffman coding method.

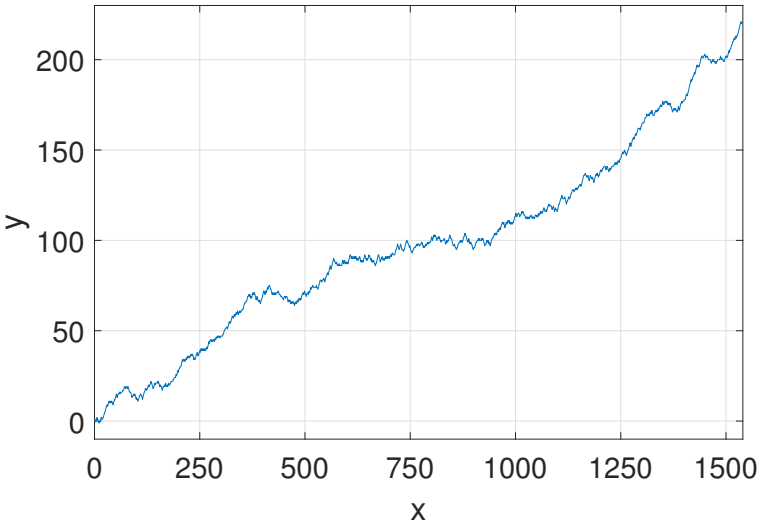

Figure 3. Corresponding graphical curves of ND5 gene for Human (H. sapiens, YP_003024036.1)

3 Numerical characterization and similarity analysis of proteins

3.1 Numerical characterization

Similarity analysis of biological sequences is an important problem in bioinformatics. There are two main directions, alignment and mathematical model. Based on the idea of alignment, there are many classical sequence comparison methods, such as BLAST2, EMBOSS, ClustalW, and so on. Based on mathematical theories researchers developed lots of mathematical methods for similarity analysis. For example, Randić et al. [11] presented E, M/M and L/L matrix to discuss similarity comparison of sequences. Other mathematical methods were also recommended to discuss the similarity problem [36–39]. These methods were useful and used to deal with more biological problems besides similarity comparison.

Here, we propose a multicomponent vector as descriptor to quantita-

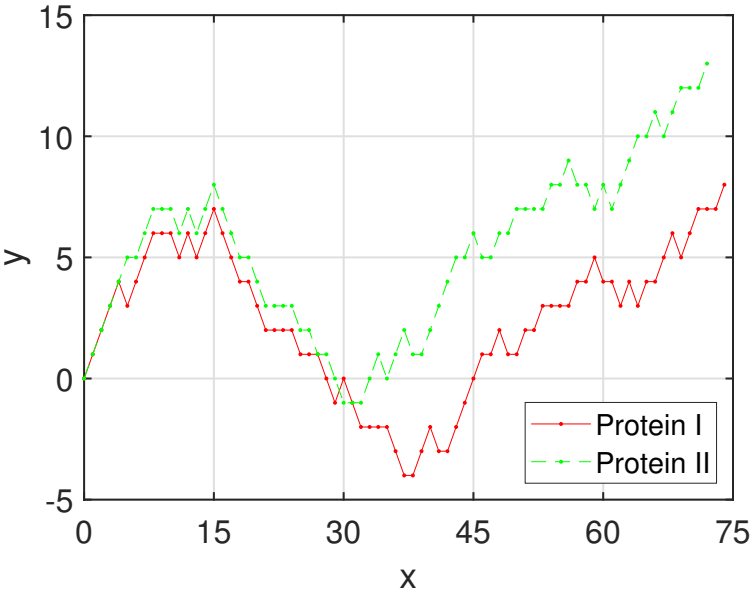

Figure 4. Graphical representation of the protein I and protein II by the proposed method

tively characterize protein sequences, and then do similarity analysis by the descriptor.

Based on the 3-ary Huffman method, a protein sequence is mapped into a digital sequence consisting of bit "0", "1" and "2". Each amino acid corresponds to a 3-ary Huffman code. Every 3-ary Huffman code of the digital sequence includes two kinds of important information, the number of the code and its position distribution. Now, let $Code_i$ be a 3-ary Huffman code of 20 amino acids, $i \in \{G, A, T, S, P, V, L, I, M, F, Y, W, D, E, N, Q, K, R, H, C\}$. The p_i is the proportion of the $Code_i$ in the 3-ary digital sequence.

$$p_i = \frac{the\ number\ of\ Code_i}{\sum the\ number\ of\ Code_i} \tag{1}$$

Let d_j be the position of the jth occurrence of the $Code_i$, $j \in \{1, 2, \cdots, k\}$, and the k is the number of the $Code_i$. All position information of the

$Code_i$ forms its position vector D_i, and $D_i = (d_1, d_2, \cdots, d_k)$. For the protein sequence of ND5 gene for Human (H. sapiens, YP_003024036.1), based on the Table 2 the 3-ary Huffman codes are the following (for simplification, several codes being listed),

$$Code_G, 200,\ Code_A, 01,\ Code_T, 12,\ Code_S, 02,\ Code_P, 210.$$

Next, we can easily get the two kinds of important information, the number of a code and its position distribution. For example, the $Code_T$, 12, its proportion is 10.95%. Its position vector is $D_T = (4, 12, 14, \cdots, 1538)$. The element of the position vector D_i denotes the position where the $Code_T$ 12 of T appears in the 0-2 digital sequence. They can effectively describe the distribution information of the $Code_i$. Due to the randomness of the distribution of a code, it is difficult to quantitatively compare the position vectors of different sequences. Therefore, the position vector needs to be normalized to a characteristic value which can effectively characterize the location feature of a given code in a sequence.

The entropy is known as a measure of the uncertainty of the probability distribution [40]. The entropy theory is extensively used to describe the degree of uncertainty in many fields. Let X be a discrete probability vector which has possible values $x_1, x_2, \cdots, x_n, \sum_{k=1}^{n} x_k = 1$. The entropy of X is defined as

$$H(X) = -\sum_{k=1}^{n} x_k \log_2 x_k. \tag{2}$$

Here, we firstly normalized the position vector $D_i = (d_1, d_2, \cdots, d_k)$. Let p_j be the jth normalized position value,

$$p_j = \frac{d_j}{\sum_{j=1}^{k} d_j}. \tag{3}$$

The parameter p_j describes the probability of the relative position j of the $Code_i$ occurrence in the 3-ary digital sequence. Next, based on the entropy theory we can define the position entropy h_i of the $Code_i$ as the

following,

$$h_i = -\sum_{j=1}^{k} p_j \log_2 p_j. \tag{4}$$

For the protein sequence (H. sapiens, YP_003024036.1), the position entropy h_i of the $Code_i$ is computed by the formula (for simplification, several entropy values being listed),

$$h_G : 4.59, h_A : 5.32, h_T : 5.65, h_S : 5.33, h_P : 4.70.$$

The most of entropy values of the codes are larger than 1. However, the proportions of the codes vary between 0 and 1. For a consistent range of variation we normalize the entropy to the range from 0 and 1 by the formula,

$$\hat{h}_i = \frac{h_i - minh}{maxh - minh}, \tag{5}$$

where $minh = \min h_j$, and $maxh = \max h_j$, $j \in \{G, A, T, S, P, V, L, I, M, F, Y, W, D, E, N, Q, K, R, H, C\}$.

The normalized results are the following,

$$\hat{h}_G : 0.54, \hat{h}_A : 0.71, \hat{h}_T : 0.80, \hat{h}_S : 0.72, \hat{h}_P : 0.56.$$

Next, we can obtain a 2-tuple $(p_i, \hat{h}_i)$ to describe the proportion and the position distribution of the $Code_i$ in its 3-ary digital sequence, $i \in \{G, A, T, S, P, V, L, I, M, F, Y, W, D, E, N, Q, K, R, H, C\}$. For the 20 amino acids, we can construct a 40-dimensional mixed characteristic vector $\vec{V}$ as follows,

$$\vec{V} = (p_G, \hat{h}_G, p_A, \hat{h}_A, p_T, \hat{h}_T, \cdots, p_C, \hat{h}_C,). \tag{6}$$

In Table 2, we have shown the 3-ary Huffman codes of the 20 amino acids of ND5 gene for Human (H. sapiens, YP_003024036.1). Based on the 2-tuple $(p_i, \hat{h}_i)$ of the codes we can obtain the 40-dimensional mixed vector $\vec{V}$ to characterize the segment.

$$\begin{aligned}\vec{V} =&(0.05, 0.54,\ 0.08, 0.71,\ 0.12, 0.80,\ 0.09, 0.72,\ 0.06, 0.56,\ 0.03, 0.28,\\ &0.18, 1.00,\ 0.10, 0.75,\ 0.05, 0.46,\ 0.07, 0.63,\ 0.03, 0.31,\ 0.02, 0.22,\\ &0.02, 0.19,\ 0.02, 0.13,\ 0.06, 0.58,\ 0.04, 0.41,\ 0.04, 0.42,\ 0.01, 0.12,\\ &0.02, 0.27,\ 0.01, 0.00).\end{aligned}$$

3.2 Similarity analysis

Similarity analysis of biological sequences is an important problem for biologists. Sensitive invariants derived from the graphic representation of DNA or protein sequence is one of the essential approaches to deal with similarity problem. For cases, some invariants, such as matrix, matrix and matrix [11], 7-vector [41], and cumulative distance [42, 43], were proposed to deal with the similarity analysis of different biological sequences.

Here, we use the 40-dimensional mixed vector $\vec{V}$ to quantitatively characterize a protein sequence, and do similarity analysis of sequences. An underlying assumption is that if two 40-dimensional mixed vectors point to a similar direction in the 40D space, the two protein sequences represented by the vectors are similar.

Let $\vec{V}_1$ and $\vec{V}_2$ be the characteristic vectors of protein sequence P_1 and P_2, respectively. The parameter α is defined as the angle between the vectors. We can calculate the cosine value of α,

$$\cos\alpha = \frac{\vec{V}_1\vec{V}_2}{|\vec{V}_1||\vec{V}_2|}. \tag{7}$$

Next, we define a distance $D(\vec{V}_1, \vec{V}_2)$ to represent the evolution relationship of protein sequence P_1 and P_2.

$$D(\vec{V}_1, \vec{V}_2) = \frac{1-\cos\alpha}{2} = \frac{1}{2}\left(1 - \frac{\vec{V}_1\vec{V}_2}{|\vec{V}_1||\vec{V}_2|}\right). \tag{8}$$

Two protein sequences are considered relatively similar if the $D(\vec{V}_1, \vec{V}_2)$ is small.

4 Applications and discussion

4.1 Comparison of nine ND6 proteins by the proposed 3-ary Huffman coding method

Table 3. The 3-ary Huffman codes of 20 amino acids within the nine DN6 proteins

G	A	T	S	P	V	L	I	M	F
10	022	020	21	1212	11	01	22	120	20
Y	W	D	E	N	Q	K	R	H	C
122	001	000	021	1211	00202	1210	0022	00201	0021

In this section, we firstly give an application for the comparison of nine DN6 proteins (Human (YP_003024037), North American opossum (NP_007 106), Gallus (NP_006927), House mouse (NP_904339), Norway rat (YP_665 640), Blue whale (NP_007067), Fin whale (NP_006900), Gorilla (NP_00822 3), and the Chimpanzee (NP_008197)). By the proposed 3-ary Huffman coding method, we construct a 3-ary Huffman tree of the nine DN6 proteins. The 0-2 codes of 20 amino acids in the 3-ary Huffman tree are shown in Table 3. By the 0-2 codes of Table 3 we can obtain the 0-2 digital sequences of the nine DN6 proteins. Next, the 0-2 digital sequences are mapped into a 40-dimensional vector set. We can compute the evolution distances of sequences by the distance formula $D(\vec{V}_1, \vec{V}_2)$.

Table 4. Similarity/dissimilarity matrix for nine ND6 proteins by the proposed method

Species	*Human*	*Opossum*	*Gallus*	*Mouse*	*Rat*	*B.whale*	*F.whale*	*Gorilla*	*Chimpanzee*
Human	0	0.01331	0.04759	0.01501	0.01698	0.01424	0.01782	0.00222	0.00053
Opossum		0	0.04926	0.01642	0.01401	0.01295	0.01323	0.01437	0.01428
Gallus			0	0.05749	0.05776	0.06127	0.06009	0.04816	0.04802
Mouse				0	0.00246	0.01932	0.02150	0.01165	0.01567
Rat					0	0.01434	0.01622	0.01500	0.01844
B.whale						0	0.00187	0.01832	0.01622
F.whale							0	0.02324	0.01975
Gorilla								0	0.00235
Chimpanzee									0

Pairwise comparison results of the nine DN6 proteins are shown in Table 4. From Table 4 we can find that Human, Gorilla and Chimpanzee have the smallest evolutionary distance. The evolutionary distance between Fin whale and Blue whale is also small. House mouse and Norway rat have a close evolutionary relationship. On the other hand, looking closely we find that North American opossum and Gallus have a further evolutionary relationship with other species.

For comparison, we make comparisons of the Human with the other eight species by the following methods, the proposed 3-ary Huffman coding method, Qi [44,45], Yao [46], He [47] EMBOSS Needle (www.ebi.ac.uk /Tools/psa/emboss_needle/nucleotide.html) and EMBOSS Water (http:// www.ebi.ac.uk/Tools/psa/emboss_water/nucleotide.html). Table 5 lists the results of the comparison. The methods in Table 5 utilize different ways, such as decimals or percentages, to describe sequence similarity. If it is a decimal, then smaller decimal means similar evolutionary distance. If it is a percentage, then larger percentage denotes similar evolutionary distance. The results of Table 5 show that the evolution distance or the similarity by the proposed 3-ary Huffman coding method is consistent with those by the other six methods although there is a slight difference in the results. Each method reveals different aspects of similarity.

Table 5. Comparisons of Human with the other eight species by different methods, the proposed 3-ary Huffman coding method, Qi [44,45], Yao [46], He [47], EMBOSS Needle and EMBOSS Water

Method	*Opossum*	*Gallus*	*Mouse*	*Rat*	*B.whale*	*F.whale*	*Gorilla*	*Chimpanzee*
This proposed mefhod	0.01331	0.04759	0.01501	0.01698	0.01424	0.01782	0.00222	0.00053
Qi [44]	0.02126	0.01148	0.00531	0.01565	0.00720	0.00802	0.00037	0.00036
Qi [45]	0.00156	—	0.00178	0.00179	0.001312	0.00126	0.00059	0.00056
Yao [46]	0.01633	0.00253	0.00770	0.01710	0.00320	0.00188	0.00033	0.00003
He [47]	275.00	—	265.70	269.38	238.91	235.46	78.81	56.43
EMBOSS (Needle)	59.1%	54.9%	71.8%	71.8%	73.1%	74.3%	97.1%	97.1%
EMBOSS (Water)	61.1%	57.7%	73.8%	73.8%	73.8%	74.7%	97.1%	97.1%

4.2 Evolution trend analysis of HA genes of the Human influenza A (H1N1) virus by the proposed method

Next, we give another application for evolutionary trend analysis of HA genes of the Human influenza A (H1N1) isolates. In April 2009, a new infectious influenza A (H1N1) virus, A/California/07/2009 (H1N1)-like virus, shows a strong ability to broadcast from human to human and spread worldwide. In [48], authors have taken into the evolution trends of the 2009 pandemic Human influenza A (H1N1) viruses from March 2009 to April 2012. When the time enters 2022, people are eager to know the recent evolutionary trend of the virus. How much mutation has occurred compared to the virus in 2009? Here, we consider two sets of HA gene data from the pandemic Human influenza A (H1N1) virus, the 401 HA gene sequences from March to April of 2009, and the 969 HA gene sequences from January of 2020 to June of 2022. The two sets of data are downloaded from the NCBI Influenza Virus Resource (http://www.ncbi.nlm.nih.gov/genomes/FLU/SwineFlu.html).

By the proposed 3-ary Huffman coding method, for every HA gene sequence we can get a corresponding 40-dimensional vector. Based on the two sets of HA gene sequences, we can obtain a vector set with 1370 40-dimensional vectors. By the proposed distance formula $D(\vec{V}_1, \vec{V}_2)$, pairwise comparison results of all vectors form a 1370×40 matrix. The matrix is too large for us to find out the evolutionary relationship of the two sets of HA gene sequences. In order to easily grasp the evolutionary characteristics, we consider their visual representation in 2D space. By the PCA method we project the 1370 40-dimensional vectors onto the 2D coordinate system of Fig.5 (Red dots, 2009; Blue dots, 2020; Black dots, 2021; Pink dots, 2022). The first and the second principal axes account for 75.3% and 11.3% of the total inertia of the 40D space, respectively. Fig.5 shows that all isolates are classified into several distinct clusters. The isolates marked by red dots were sampled from March to April of 2009 and grouped into two distinct clusters. This indicates that a new mutant strain has appeared from March to April of 2009. The evolutionary fact is that in April 2009

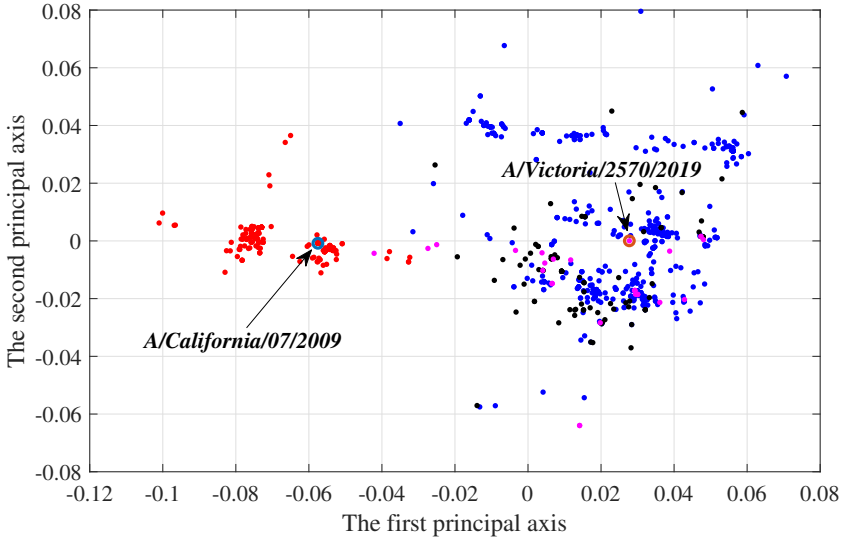

Figure 5. The 2D graphical mapping of HA genes of 1370 influenza A (H1N1) isolates from March to April of 2009 and January of 2020 to June of 2022 (Red dots, 2009; Blue dots, 2020; Black dots, 2021; Pink dots, 2022). The first and second principal axes account for 75.3% and 11.3% of the total inertia of the 20D space, respectively.

a new mutant strain, A/California/07/2009 (H1N1)-like virus, emerged and had a strong ability to broadcast from human to human and spread worldwide. Fig.5 shows the new mutant cluster with the strain, A/California/07/2009 (H1N1). The strain became the recommended vaccine strain for H1N1 virus in the following years. The results are also consistent with the works [44, 48].

In order to explore the recent evolutionary trend of the virus, in Fig.5 we show the other data set from January 2020 to June 2022. Taking a close observing, we find that there are three very obvious characteristics. One is that the virus population since January 2020 has been significantly far away from that in 2009. The second is that in recent years, the virus populations have fused with each other. By World Health Organization (WHO) the recommended vaccine strain for H1N1 virus since January 2020 has been the strain, A/Victoria/2570/2019 (H1N1). In Fig.5 we also

show the recommended vaccine strain.

The third characteristic is that in recent years, the virus population has been gradually moving away from the vaccine strain, A/Victoria/2570/2019 (H1N1), in a way of diffusion, although they are still around the vaccine strain. This deviation from the vaccine strain seems to be getting faster and faster. The high deviation speed means that the virus population are likely experiencing great selection pressure. The virus genome would become no longer stable. The results emphasize the importance of continuous monitoring of the 2009 pandemic influenza A (H1N1) in the following years. It is also crucial to carefully monitor the underlying evolutionary changes of the virus population.

4.3 Evolution analysis of 95 coronavirus genes by the proposed method

On 30 January 2020, the World Health Organization (WHO) declared that the outbreak of the novel coronavirus (2019-nCoV) has been a public health emergency of international concern. Novel coronavirus (2019-nCoV) showed strong human-to-human transmission ability [49]. So far it has caused a large global outbreak. As for another application, this paper gives the evolutionary analysis of 95 coronavirus gene sequences to discover the evolutionary relationship between 89 COVID-19 gene sequences and other 6 coronaviruses from different species. The data set includes two Pangolin samples (EPI_ISL_410538 and EPI_ISL_410721), one Rhinolophus affinis sample (BetaCoV/bat/Yunnan/RaTG13/2013 (EPI_ISL_402131)) from GISAID, two SARS-related coronaviruses samples (MG772933 and MG772934) from NCBI, the first shared human COVID-19 sample (BetaCoV/Wuhan-Hu-1/2019 (EPI_ISL_402125)) from China, and other 89 human COVID-19 samples from GenBank.

Based on the 3-ary Huffman coding method, the 95 gene sequences are converted into 95 corresponding 40-dimensional vectors. Next, we compute the evolution distances of sequences by the distance formula $D(\vec{V_1}, \vec{V_2})$. These calculation results are used to perform phylogenetic tree analysis by the Molecular Evolutionary Genetics Analysis (MEGA-

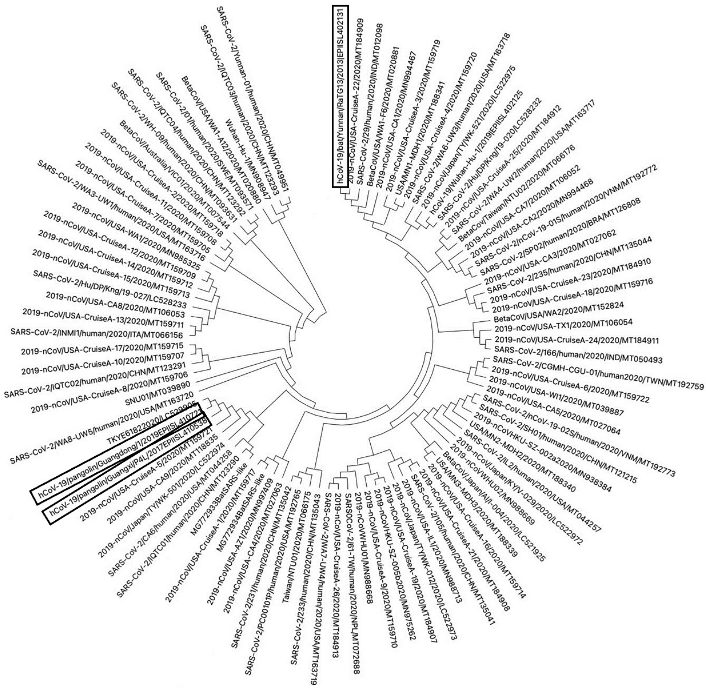

Figure 6. Phylogenetic tree of 95 coronaviruses by the proposed method

X) software [50]. The phylogenetic tree of the 95 coronaviruses samples is shown in Fig. 6. Fig. 6 shows that the Rhinolophus affinis sample (EPI_ISL_402131) is more similar to the human samples than the two Pangolin samples (EPI_ISL_410538 and EPI_ISL_410721). The results are consistent with those in Li et al. [51] and Qi [45].

5 Conclusion

This paper gives a new digital mapping method of protein sequence based on 3-ary Huffman coding algorithm. The mapping is unique when a 3-ary Huffman tree is defined by the frequency characteristic of 20 amino acids in given protein sequences. The unique 0-2 digital sequence of a protein sequence can be used to perform graphical representation of the protein

sequence and sequence feature extraction. Next, we construct multicomponent vectors based on the frequency characteristic and the distribution information of 0-2 codes of 20 amino acids in the 0-2 digital sequence. The vector is used to calculate the similarity distance between two protein sequences. To illustrate the effectiveness of the method, we apply it to three separate applications, similarity comparison of nine ND6 proteins, evolutionary trend analysis of the 2009 pandemic Human influenza A (H1N1) viruses from January 2020 to June 2022, and the evolution analysis of 95 coronavirus genes. The results show the utility of the proposed method.

With the discovery of the 21st and 22nd amino acids (Selenocysteine and Pyrrolysine), some rare protein sequences may contain the two amino acids. The proposed method can also be easily extended to deal with protein sequences containing 22 amino acids. The only difference is that the 3-ary Huffman tree is defined by the frequency characteristic of 22 amino acids in given protein sequences. The 0-2 codes of 22 amino acids constructed by the 3-ary Huffman tree can convert long protein sequences into one-to-one 0-2 digital sequences. The other processes are the same as considering 20 amino acids.

***Acknowledgment*:** This work is supported by Humanities and Social Sciences Research of Ministry of Education of China (Grant No. 19YJAZH0 69) and Hunan Provincial Science and Technology Project Foundation (Grant No. 2018TP1018)

The authors declare that they have no conflict of interest.

References

[1] T. D. Pham, J. Zuegg, A probabilistic measure for alignment-free sequence comparison, *Bioinformatics* **20** (2004) 3455–3461.

[2] S. Vinga, J. Almeida, Alignment-free sequence comparison—a review, *Bioinformatics* **19** (2003) 513–523.

[3] E. Hamori, J. Ruskin, H curves, A novel method of representation of nucleotide series especially suited for long DNA sequences, *J. Biol. Chem.* **258** (1983) 1318–1327.

[4] E. Hamori, Graphic representation of long DNA sequences by the method of Hcurves-current results and future aspects, *BioTechniques* **7** (1989) 710–720.

[5] H. J. Jeffrey, Chaos game representation of gene structure, *Nucleic Acids Res.* **18** (1990) 2163–2170.

[6] A. Nandy, A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes, *Curr. Sci.* **66** (1994) 309–314.

[7] D. Bielińska-Wa ż, Four-component spectral representation of DNA sequences, *J. Math. Chem.* **47** (2010) 41–51.

[8] D. Bielińska-Wa ż, W. Nowak, P. Wa ż, A. Nandy, T. Clark, Distribution moments of 2D-graphs as descriptors of DNA sequences, *Chem. Phys. Lett.* **443** (2007) 408–413.

[9] B. Liao, C. Zeng, F. Li, Y. Tang, Analysis of similarity/dissimilarity of DNA sequences based on dual nucleotides, *MATCH Commun. Math. Comput. Chem.* **59** (2008) 647–652.

[10] B. Liao, Q. Xiang, L. Cai, Z. Cao, A new graphical coding of DNA sequence and its similarity calculation, *Phys. A* **392** (2013) 4663–4667.

[11] M. Randić, M. Vračko, N. Lerš, D. Plavšić., Novel 2-D graphical representation of DNA sequences and their numerical characterization, *Chem. Phys. Lett.* **368** (2003) 1–6.

[12] M. Randić, Another look at the chaos-game representation of DNA, *Chem. Phys. Lett.* **456** (2008) 84–88.

[13] G. Jaklič, T. Pisanski, M. Randić, Characterization of complex biological systems by matrix invariants, *J. Comput. Biol.* **13** (2006) 1558–1564.

[14] Z. Qi, L. Li, X. Qi, Using Huffman coding method to visualize and analyze DNA sequences, *J. Comput. Chem.* **32** (2011) 3233–3240.

[15] M. Randić, 2-D graphical representation of proteins based on virtual genetic code, *SAR QSAR Environ. Res.* **15** (2004) 147–157.

[16] M. Randić, J. Zupan, A. T. Balaban, Unique graphical representation of protein sequences based on nucleotide triplet codons, *Chem. Phys. Lett.* **397** (2004) 247–252.

[17] M. Randić, D. Butina, J. Zupan, Novel 2-D graphical representation of proteins, *Chem. Phys. Lett.* **419** (2006) 528–532.

[18] F. Bai, T. Wang, A 2-D graphical representation of protein sequences based on nucleotide triplet codons, *Chem. Phys. Lett.* **413** (2005) 458–462.

[19] P. He, D. Li, Y. Zhang, X. Wang, Y. Yao, A 3D graphical representation of protein sequences based on the Gray code, *J. Theor. Biol.* **304** (2012) 81–87.

[20] M. Randić, 2-D graphical representation of proteins based on physicochemical properties of amino acids, *Chem. Phys. Lett.* **444** (2007) 176–180.

[21] Y. Yao, Q. Dai, C. Li, P. He, X. Nan, Y. Zhang, Analysis of similarity/dissimilarity of protein sequences, *Proteins: Struct. Func. Bioinf.* **73** (2008) 864–871.

[22] M. I. A. el Maaty, M. M. Abo-Elkhier, M. A. Abd Elwahaab, 3D graphical representation of protein sequences and their statistical characterization, *Phys. A* **389** (2010) 4668–4676.

[23] Z. Wu, X. Xiao, K. Chou, 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** (2010) 29–34.

[24] B. E. Blaisdell, A measure of the similarity of sets of sequences not requiring sequence alignment, *Proc. Natl. Acad. Sci.* **83** (1986) 5155–5159.

[25] G. W. Stuart, K. Moffett, S. Baker, Integrated gene and species phylogenies from unaligned whole genome protein sequences, *Bioinformatics* **18** (2002) 100–108.

[26] T. J. Wu, J. P. Burke, D. B. Davison, A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words, *Biometrics* **53** (1997) 1431–1439.

[27] V. Afreixo, C. A. Bastos, A. J. Pinho, S. P. Garcia, P. J. Ferreira, Genome analysis with inter-nucleotide distances. *Bioinformatics* **25** (2009) 3064–3070.

[28] Y. Gao, L. Luo, Genome-based phylogeny of dsDNA viruses by a novel alignment-free method, *Gene* **492** (2012) 309–314.

[29] S. Ding, Y. Li, X. Yang, T. Wang, A simple k-word interval method for phylogenetic analysis of DNA sequences, *J. Theor. Biol.* **317** (2013) 192–199.

[30] Q. Dai, X Liu, Y. Yao, F. Zhao, Numerical characteristics of word frequencies and their application to dissimilarity measure for sequence comparison, *J. Theor. Biol.* **276** (2011) 174–180.

[31] L. Yang, X. Zhang, H. Zhu, Alignment free comparison: similarity distribution between the DNA primary sequences based on the shortest absent word, *J. Theor. Biol.* **295** (2012) 125–131.

[32] D. Huang, H. Yu, Normalized feature vectors: a novel alignment-free sequence comparison method based on the numbers of adjacent amino acids, *IEEE/ACM Trans. Comput. Biol. Bioinf.* **10** (2013) 457–467.

[33] F. Bai, J. Xu, L. Liu, Weighted relative entropy for phylogenetic tree based on 2-step Markov model, *Math. Biosci.* **246** (2013) 8–13.

[34] Z. Qi, M. Jin, J. Wang, S. Li, Novel DNA sequence comparison method based on Markov chain and information entropy, *Mini Rev. Org. Chem.* **12** (2015) 524–533.

[35] D. A. Huffman, A method for the construction of minimum-redundancy codes, *Proc. IRE.* **40** (1952) 1098–1102.

[36] Z. Qi, T. Fan, PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization, *Chem. Phys. Lett.* **442** (2007) 434–440.

[37] J. Yu, X. Sun, J. Wang, TN curve: a novel 3D graphical representation of DNA sequence based on trinucleotides and its applications, *J. Theor. Biol.* **261** (2009) 459–468.

[38] J. Yu, X. Sun, Reannotation of protein-coding genes based on an improved graphical representation of DNA sequence, *J. Comput. Chem.* **31** (2010) 2126–2135.

[39] D. Panas, P. Wa ż, D. Bielińska-Wa ż, A. Nandy, S. C. Basak, An application of the 2D-dynamic representation of DNA/RNA sequences to the prediction of influenza A virus subtypes, *MATCH Commun. Math. Comput. Chem.* **80** (2018) 295–310.

[40] D. J. MacKay,D. J. Mac Kay, *Information Theory, Inference, and Learning Algorithms*, Cambridge Univ. Press, Cambridge, 2003.

[41] Z. Qi, X. Qi, Novel 2D graphical representation of DNA sequence based on dual nucleotides, *Chem. Phys. Lett.* **440** (2007) 139–144.

[42] M. Randić, K. Mehulić, D. Vukičević, T. Pisanski, D. Vikić-Topić, D. Plavšić, Graphical representation of proteins as four-color maps and their numerical characterization, *J. Mol. Graphics Modell.* **27** (2009) 637–641.

[43] B. Liao, B. Liao, X. Sun, Q. Zeng, A novel method for similarity analysis and protein sub-cellular localization prediction, *Bioinformatics* **26** (2010) 2678–2683.

[44] Z. Qi, M. Jin, S. Li, F. Jun, A protein mapping method based on physicochemical properties and dimension reduction, *Comput. Biol. Med.* **57** (2015) 1–7.

[45] Z. Qi, X. Wen, Novel protein sequence comparison method based on transition probability graph and information entropy, *Chem. High Throughput Screen.* **25** (2022) 392–400.

[46] Y. Yao, S. Yan, J. Han, Q. Dai, P. A. He, A novel descriptor of protein sequences and its application, *J. Theor. Biol.* **347** (2014) 109–117.

[47] C. Li, Q. Dai, P. A. He, A time series representation of protein sequences for similarity comparison, *J. Theor. Biol.* **538** (2022) #111039.

[48] Z. Qi, J. Feng, C. Liu, Evolution trends of the 2009 pandemic influenza A (H1N1) viruses in different continents from March 2009 to April 2012, *Biologia* **69** (2014) 407–418.

[49] J. F. W. Chan, S. Yuan, K. H. Kok, K. K. W. To, H. Chu, J. Yang, F. Xing, J. Liu, C. C. Y. Yip, R. W. S. Poon, H. W. Tsoi, S. K. Lo, K. H. Chan, V. K. Poon, W. M. Chan, J. D. Ip, J. P. Cai, V. C. Cheng, H. Chen, C. K. Hui, K. Y. Yuen, A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster, *Lancet* **395** (2020) 514–523.

[50] S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: molecular evolutionary genetics analysis across computing platforms, *Mol. Biol. Evol.* **35** (2018) 1547–1549.

[51] X. Li, J. Zai, Q. Zhao, Q. Nie, Y. Li, B. T. Foley, A. Chaillon, Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2, *J. Med. Virol.* **92** (2020) 602–611.