Communications in Information and Systems
Volume 22, Number 3, 317–337, 2022

Nucleotide amino acid k-mer vector: an alignment-free method for comparing genomic sequences

Xiaona Bao[*], Lily He[*], Jingan Cui[†], and Stephen S.-T. Yau[†]

Dedicate to professor Michael Waterman on his 80th birthday

Evolutionary analysis of genomic data is a valuable issue in the study of bioinformatics, and a great deal of DNA data has become available. In the field of evolutionary analysis, protein sequences are more meaningful than DNA sequences, and the alignment-free methods based on k-mer mean are widely used. However, the dimension of the k-mer vector based on protein sequence is very high. This paper proposes a new Nucleotide Amino Acid K-mer Vector (NAAKV) technique, which converts the DNA sequence to a pseudo amino acid sequence (PAAS). This transformation does not need to find the coding region of the gene sequence, but also reflects the change of nucleotide. Meanwhile, there is a strong correlation between the amino acids, which leads to the types of k-mer are much lower than that of protein sequence, thus the dimension is greatly reduced. To test NAAKV, we carry out phylogenetic analysis of several viruses and bacteria. The traditional k-mer method and alignment-based MUSCLE method are used for comparison on each dataset. Eventually, the results suggest that NAAKV is accurate and time-efficient for phylogenetic analysis and genome classification.

AMS 2000 subject classifications: Primary 68W50, 92D20; secondary 03D32.
Keywords and phrases: Alignment-free, k-mer, NAAKV, evolutionary analysis, DNA sequence.

1. Introduction

The progress of sequencing technology has made a vast amount of biological data accessible. Moreover, sequence comparison and evolutionary analysis

[*] These authors contribute equally to this work and should be considered co-first authors.

[†] Co-Corresponding authors.

of DNA sequences are an important area in biology. This technique can help treat diseases caused by specific bacteria or viral strains more effectively. For example, the recent pandemic COVID-19 is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1, 2]. SARS-CoV-2 is a new strain of coronavirus that has never been found in the human body before. The virus has produced a variety of different pathogenic types [3, 4]. It is threatening human health all over the world. Cervical cancer that caused by human papillomavirus (HPV) is also a serious high-risk disease among women. Moreover, there are many subtypes of HPV that can cause diseases with different symptoms [5]. Therefore, effective genome classification methods are very necessary. The traditional method is Multiple Sequence Alignment (MSA). There are a lot of software tools implement these methods for sequence comparisons, such as BLAST [6], ClustalW [7] and MUSCLE [8]. However, multiple sequence alignments (MSA) take a long time in genome comparison. So, it is difficult for them to analyze numerous sequences efficiently [9]. Therefore, researchers have begun to explore alignment-free techniques, from which several new methods have been developed [10, 11, 12, 13, 14]. These techniques make use of the length of common substrings [10], the positions of nucleotides [14], and words or k-mer counts [15, 16, 17]. Note that many alignment-free approaches involve k-mer counting. Specifically, the traditional k-mer method has obtained a lot of success in sequence comparison [18]. FFP calculates the distance between nucleotide or amino acid sequences by calculating the frequency of different k-mer types and dividing this total by the total count of all k-mers [16, 17]. The distribution of k-mer intervals is proposed to calculate the distance between genome sequences [19]. ALFRED-G employs an algorithm to calculate the length of maximal k-mismatch common substrings between nucleotide or amino acid sequences [20]. RTD-Phylogeny calculates the distances from nucleotide or protein sequences by calculating the time of k-mer reproduction [21]. KWIP estimates the genetic difference between next-generation sequencing data [13]. In general, based on distance among k-mer frequencies vectors from raw nucleotide or protein sequence data, we can construct phylogenetic trees for species [22]. But when doing genome comparison based on protein sequences, such as FFP, ALFRED-G, and RTD-Phylogeny, the dimension of k-mer vector is very high.

In this paper, we transform a given nucleotide sequence into a pseudo amino acid sequence (PAAS). During this implementation process, we translate codons into amino acids. This is not a biological translation process because it follows a given order; that is, one base is moved at a time, not three. After obtaining the pseudo amino acid sequence (PAAS), we count its k-mer types and then construct a nucleotide amino acid k-mer vector (NAAKV), consisting of the calculated k-mer frequency. Finally, the distance between different genomes is calculated by computing their vectors. PAAS contains relatively fewer k-mer types than protein sequence because many k-mer kinds do not appear in PAAS. For example, the amino acid combination "HQR" does not exist. In this way, our vector has a lower dimension and takes up less memory. This paper verifies NAAKV by analyzing complete genomes from SARS-CoV-2, bacteria, HCV, HRV and HBV. The NAAKV method conducts an accurate evolutionary analysis for each dataset. We also use these genomic datasets to compare our method with the traditional k-mer method, protein sequence and MUSCLE, revealing that NAAKV has obvious advantages concerning time and accuracy.

2. Methods

2.1. Pseudo amino acid sequence (PAAS)

Let $S = s_1 s_2, \cdots, s_n$ ($s_i \in \{A,\ T,\ C,\ G\}$, $i = 1, 2, \cdots, n$) be a DNA sequence, the process is as follows:

Starting from the first nucleotide to translate the first amino acid, then move one base at a time and start coding the next codon from here (at this time, three consecutive nucleotides constitute a codon). To be specific, $s_1 s_2 s_3$ is translated as the corresponding amino acid p_1, $s_2 s_3 s_4$ is translated as p_2 and so on, such that $s_i s_{i+1} s_{i+2}$ is always translated as the corresponding amino acid p_i. Follow this rule in turn, we can get the pseudo amino acid sequence (PAAS): $P_S = p_1 p_2, \cdots, p_{n-2}$, where $p_j \in \{A,\ C,\ D,\ E,\ F,\ G,\ H,\ I,\ K,\ L,\ M,\ N,\ P,\ Q,\ R,\ S,\ T,\ V,\ W,\ Y\}$. Obviously, the length of P_S is 2 lengths shorter than the DNA sequence S.

For example: Given a genomic sequence: S=CGATCAGCAAGC (the length of S is 12), three consecutive bases correspond to amino acids (aa) as follows:

Position:	1	2	3	4	5	6	7	8	9	10	11	12
Sequence:	**C**	**G**	**A**	T	C	A	G	C	A	A	G	C
		R (the first aa)										
Position:	1	2	3	4	5	6	7	8	9	10	11	12
Sequence:		**G**	**A**	**T**	C	A	G	C	A	A	G	C
			D (the second aa)									
Position:	1	2	3	4	5	6	7	8	9	10	11	12
Sequence:			**A**	**T**	**C**	A	G	C	A	A	G	C
				I (the third aa)								

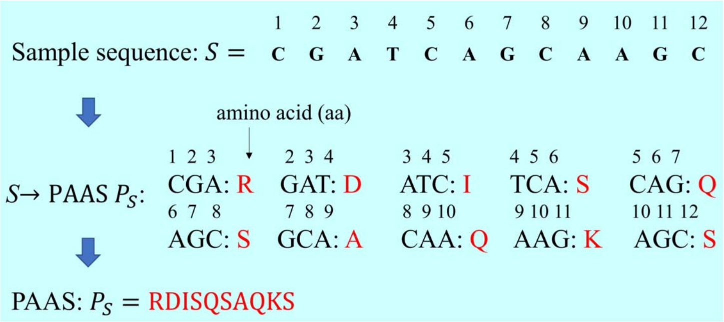

Figure 1: The flow chart of NAAKV method using a short sequence as an example. R, Arginine; D, Aspartic acid; I, Isoleucine; S, Serine; Q, Glutamine; A, Alanine; K, Lysine.

...

Position:	1	2	3	4	5	6	7	8	9	10	11	12
Sequence:										**A**	**G**	**C**
											S (the last aa)	

Details: CGA: R; GAT: D; ATC: I; TCA: S; CAG: Q; AGC: S; GCA: A; CAA: Q; AAG: K; AGC: S. Then, the PAAS P_S=RDISQSAQKS (the length of P_S is 10) is obtained, as shown in Fig. 1.

For a nucleotide sequence, by moving only one nucleotide, each triplet overlap with the next one, which leads to the translated two amino acids correlate strongly. This property is able to reflect the correlation between nucleotides. Using this particular conversion rule, we generate the PAAS whose amino acid has a relatively strong correlation. While moving the natural 3 nucleotides, the translated two amino acids seem independent from each other. Thus, the resulted natural amino acid sequence cannot capture the correlation in nucleotide sequences as well as our method.

2.2. K-mer types of pseudo amino acid sequence

K-mer method is used to count the frequency of all possible substrings of length k in DNA sequence [18]. Here, one interesting finding is that the total number of k-mer types in PAAS are less than those of protein sequence (PS). The reason is that a majority of possible cases will never appear in PAAS. Take threonine (T) and its subsequent amino acids as an example: As shown in Fig. 2, four codons (ACT, ACC, ACA, ACG) can encode threonine (T), according to the principle of PAAS, the next codons after T may have

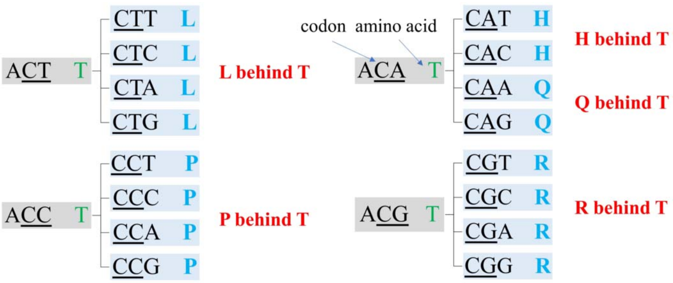

Figure 2: According to the NAAKV process, the schematic diagram of amino acid types which can appear after Threonine (T). All codons are shown in black. Threonine (T) and amino acids after T are shown in green and blue, respectively.

Table 1: All the possible 95 kinds of 2-mer

2-mer	Number
AH AL AP AQ AR CA CV DI DM DT	10
EK EN ER ES FF FL FS GA GD GE GG GV	12
HI HM HT IF IK IL IN IR IS IY	10
KK KN KR KS LC LD LE LF LK LL LN LR LS LW LY	15
MC MD ME MW NI NM NT	7
PH PL PP PQ PR QK QN QR QS	9
RA RE RD RG RV SA SH SL SP SQ SR SV	12
TH TL TP TQ TR WG YI YM YT	9
VC VD VE VF VK VL VN VR VS VW VY	11
	total number: 95

4*4=16 types. However, different codons may encode the same amino acid, so there are only 5 types, which is less than 16.

According to the rules of codon translation into amino acids, we can compute all the possible k-mer types based on PAAS. Take k=2 as an example, we get all possible 95 kinds of 2-mers as shown in Table 1. Note that the protein sequence may produce 400 kinds of 2-mers. The results of k=3 and 4 are are shown in supplementary tables http://intlpress.com/site/pub/files/supp/cis/2022/0022/0003/cis-2022-0022-0003-s001.zip.

For k=1, 2, 3, 4, we also figure out the total number of all available k-mers. The result is show in Table 2. Mention that there are 20^k potential k-mer types for a protein sequence. Thus, our k-mer method based on PAAS clearly reduces the well-known k-mer method based on natural sequences and then greatly saves the calculation time for k-mer frequency.

Table 2: The total number of all possible k-mer types of pseudo amino acid sequences (PAAS) and protein sequence (PS)

	k=1	k=2	k=3	k=4
PAAS	20	95	448	2015
PS	20	$20^2 = 400$	$20^3 = 8000$	$20^4 = 160{,}000$

2.3. Nucleotide amino acid k-mer vector (NAAKV)

For a DNA sequence $S = s_1 s_2, \cdots, s_n$ ($s_i \in \{A,\ T,\ C,\ G\}$, $i = 1, 2, \cdots, n$), we first obtain its PAAS: P_S; then the frequency of all possible k-mer types appearing in the P_S are counted, and $f_{p_i p_j}$ denotes as the frequency of p_i and p_j (p_i and $p_j \in \{A,\ C,\ D,\ E,\ F,\ G,\ H,\ I,\ K,\ L,\ M,\ N,\ P,\ Q,\ R,\ S,\ T,\ V,\ W,\ Y\}$); finally, according to the order of the alphabet, all the frequencies are put together to form a nucleotide amino acid k-mer vector (NAAKV). For example, with k=2, the NAAKV is as follows:

$$(f_{AH}, f_{AL}, f_{AP}, f_{AQ}, f_{AR}, \ldots,)$$

For any k, the k-mer vector is denoted as:

$$(f_1, f_2, ..., f_N),$$

where N is the total number of all possible k-mers.

In this study, for a given dataset, the value of k is belong to $[\lfloor \log_{20} min \rfloor - 2, \lfloor \log_{20} min \rfloor + 2]$, where $\lfloor \cdot \rfloor$ is the rounding down of $\cdot$, $[\cdot]$ stands for the closed interval of $\cdot$, and min is the minimum length of all sequences in the dataset.

Given a dataset which includes N DNA sequences, N vectors can be gotten by using the NAAKV method. Next, Euclidean distance is utilized to compute the distance matrix. Finally, a phylogenetic tree will be built via using the neighbor-joining algorithm with Mega X [23]. All calculations in this article are performed on a laptop. This computer is equipped with a Windows 10 system, and the processor is Intel Core i5-8250U CPU @ 1.60 GHz and 8GB RAM, together with python 3.8. The python source code and data in this paper are freely available to the public upon request.

3. Results

We use five datasets to test our NAAKV method. The basic information of them is shown in Table 3. The SARS-CoV-2 datasets are attained from GISAID (https://www.gisaid.org/). The Bacteria, Hepatitis C virus (HCV),

Table 3: Summary of datasets, including the number of genomes, minimum length, median length and maximum length

Dataset	Genome number	Min (bp)	Median (bp)	Max (bp)
SARS-CoV-2	127	29,562	29,805	30,566
Bacteria	59	844,775	4,016,947	5,966,919
HCV	82	8,957	9,442	9,666
HRV	116	6,944	7,135.5	7,358
HBV	151	3,182	3,215	3,248

Human rhinovirus (HRV) and Hepatitis B virus (HBV) datasets are downloaded from NCBI (https://www.ncbi.nlm.nih.gov/). For comparison, the traditional k-mer method and a typical multiple sequence alignment method MUSCLE are used. The k value of the traditional k-mer method is set to be $[\lfloor \log_4 min \rfloor, \lceil \log_4 max \rceil]$, where $\lceil \cdot \rceil$ is the rounding up of $\cdot$, min and max is the minimum and maximum length of all sequences in a dataset, respectively. Table 4 shows the k value and running time of these techniques.

Table 4: K value and operation time of the NAAKV, k-mer and MUSCLE methods. "s": second; "min": minute; "$\sim$": cannot be calculated with the device used in this article

Dataset	NAAKV	Running time	K-mer	Running time	MUSCLE
SARS-CoV-2	4	14.97 s	7	77.02 s	$\sim$
Bacteria	2	272.18 s	9	30.62 min	$\sim$
HCV	4	18.72 s	6	43.88 s	$>$ 30 min
HRV	2	1.72 s	6	57.19 s	$>$ 30 min

3.1. SARS-CoV-2

The outbreak of the coronavirus disease 2019 (COVID-19) happened in Wuhan, Hubei province, China in late 2019, and spread rapidly all over the world [2]. The etiology of COVID-19 is a positive single-stranded RNA virus SARS-CoV-2. People infected with the virus have different degrees of symptoms, including fever, cough, pneumonia and even death in severe cases. The virus has multiple routes of transmission, such as respiratory droplets and contact transmission. Besides, SARS-CoV-2 mutate rapidly. So, it is difficult to prevent and control COVID-19.

GISAID website gives the clades of SARS-CoV-2. Here, we take a sum of 127 complete genomes of ten clades, namely S, O, L, V, G, GH, GK, GR, GV and GRY. The phylogenetic trees constructed from these sequences by

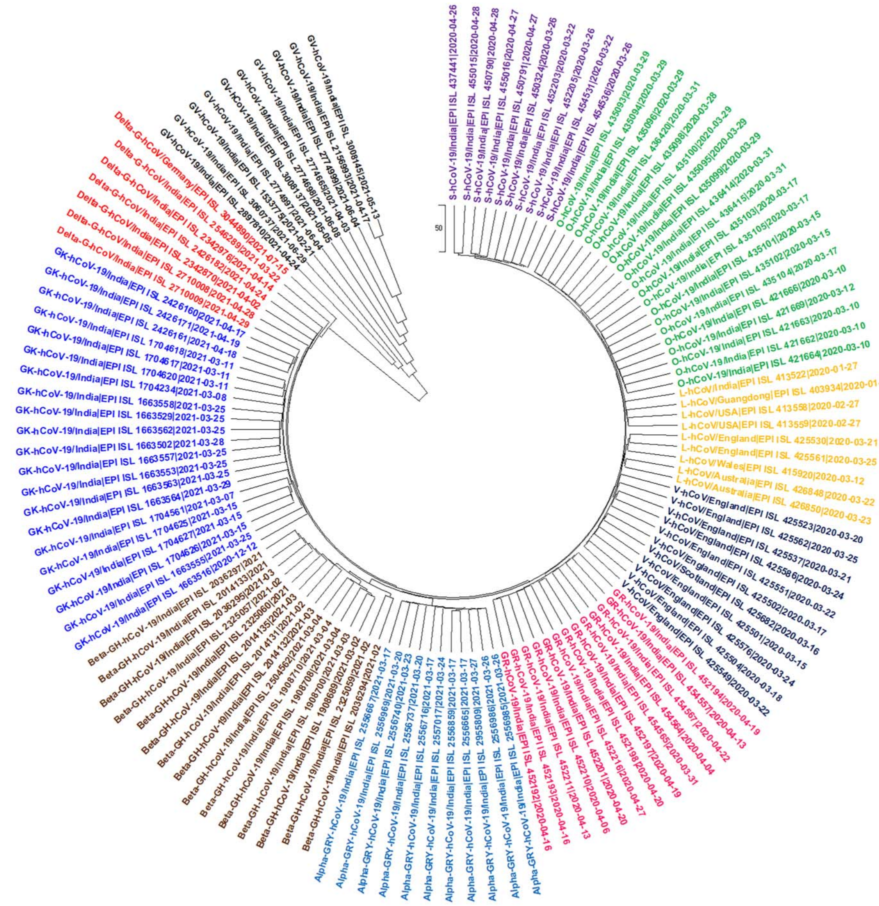
50
S-hCoV-19/India|EPI ISL 437441|2020-04-26
S-hCoV-19/India|EPI ISL 455015|2020-04-28
S-hCoV-19/India|EPI ISL 450790|2020-04-28
S-hCoV-19/India|EPI ISL 455016|2020-04-27
S-hCoV-19/India|EPI ISL 450791|2020-04-27
S-hCoV-19/India|EPI ISL 450324|2020-03-26
S-hCoV-19/India|EPI ISL 452203|2020-03-22
S-hCoV-19/India|EPI ISL 452205|2020-03-26
S-hCoV-19/India|EPI ISL 454531|2020-03-22
S-hCoV-19/India|EPI ISL 454536|2020-03-26
O-hCoV-19/India|EPI ISL 435093|2020-03-29
O-hCoV-19/India|EPI ISL 435094|2020-03-29
O-hCoV-19/India|EPI ISL 435096|2020-03-29
O-hCoV-19/India|EPI ISL 436420|2020-03-31
O-hCoV-19/India|EPI ISL 435098|2020-03-28
O-hCoV-19/India|EPI ISL 435100|2020-03-29
O-hCoV-19/India|EPI ISL 435095|2020-03-29
O-hCoV-19/India|EPI ISL 435099|2020-03-29
O-hCoV-19/India|EPI ISL 436414|2020-03-31
O-hCoV-19/India|EPI ISL 436415|2020-03-31
O-hCoV-19/India|EPI ISL 435103|2020-03-17
O-hCoV-19/India|EPI ISL 435105|2020-03-17
O-hCoV-19/India|EPI ISL 435101|2020-03-15
O-hCoV-19/India|EPI ISL 435102|2020-03-15
O-hCoV-19/India|EPI ISL 435104|2020-03-17
O-hCoV-19/India|EPI ISL 421666|2020-03-10
O-hCoV-19/India|EPI ISL 421669|2020-03-12
O-hCoV-19/India|EPI ISL 421663|2020-03-10
O-hCoV-19/India|EPI ISL 421662|2020-03-10
O-hCoV-19/India|EPI ISL 421664|2020-03-10
L-hCoV/India|EPI ISL 413522|2020-01-27
L-hCoV/Guangdong|EPI ISL 403934|2020-01-15
L-hCoV/USA|EPI ISL 413558|2020-02-27
L-hCoV/USA|EPI ISL 413559|2020-02-27
L-hCoV/England|EPI ISL 425530|2020-03-21
L-hCoV/England|EPI ISL 425561|2020-03-25
L-hCoV/Wales|EPI ISL 415920|2020-03-12
L-hCoV/Australia|EPI ISL 426848|2020-03-22
L-hCoV/Australia|EPI ISL 426850|2020-03-23
V-hCoV/England|EPI ISL 425523|2020-03-20
V-hCoV/England|EPI ISL 425562|2020-03-25
V-hCoV/England|EPI ISL 425537|2020-03-21
V-hCoV/England|EPI ISL 425586|2020-03-24
V-hCoV/England|EPI ISL 425551|2020-03-22
V-hCoV/Scotland|EPI ISL 425502|2020-03-17
V-hCoV/England|EPI ISL 425682|2020-03-16
V-hCoV/England|EPI ISL 425501|2020-03-15
V-hCoV/England|EPI ISL 425576|2020-03-24
V-hCoV/England|EPI ISL 425504|2020-03-18
V-hCoV/England|EPI ISL 425549|2020-03-22
GR-hCoV-19/India|EPI ISL 452194|2020-04-19
GR-hCoV-19/India|EPI ISL 454557|2020-04-13
GR-hCoV-19/India|EPI ISL 454567|2020-04-22
GR-hCoV-19/India|EPI ISL 454564|2020-04-04
GR-hCoV-19/India|EPI ISL 454565|2020-03-31
GR-hCoV-19/India|EPI ISL 452197|2020-04-19
GR-hCoV-19/India|EPI ISL 452198|2020-04-20
GR-hCoV-19/India|EPI ISL 452216|2020-04-27
GR-hCoV-19/India|EPI ISL 452201|2020-04-20
GR-hCoV-19/India|EPI ISL 452210|2020-04-06
GR-hCoV-19/India|EPI ISL 452211|2020-04-13
GR-hCoV-19/India|EPI ISL 452193|2020-04-16
GR-hCoV-19/India|EPI ISL 452192|2020-04-16
Alpha-GRY-hCoV-19/India|EPI ISL 2556985|2021-03-26
Alpha-GRY-hCoV-19/India|EPI ISL 2556986|2021-03-26
Alpha-GRY-hCoV-19/India|EPI ISL 2955809|2021-03-27
Alpha-GRY-hCoV-19/India|EPI ISL 2556665|2021-03-17
Alpha-GRY-hCoV-19/India|EPI ISL 2556859|2021-03-17
Alpha-GRY-hCoV-19/India|EPI ISL 2557017|2021-03-24
Alpha-GRY-hCoV-19/India|EPI ISL 2556716|2021-03-17
Alpha-GRY-hCoV-19/India|EPI ISL 2556737|2021-03-20
Alpha-GRY-hCoV-19/India|EPI ISL 2556740|2021-03-23
Alpha-GRY-hCoV-19/India|EPI ISL 2556969|2021-03-20
Alpha-GRY-hCoV-19/India|EPI ISL 2556667|2021-03-17
Beta-GH-hCoV-19/India|EPI ISL 2036294|2021-02
Beta-GH-hCoV-19/India|EPI ISL 2325059|2021-02
Beta-GH-hCoV-19/India|EPI ISL 1908689|2021-03-02
Beta-GH-hCoV-19/India|EPI ISL 1908700|2021-03-03
Beta-GH-hCoV-19/India|EPI ISL 1908708|2021-03-04
Beta-GH-hCoV-19/India|EPI ISL 1908710|2021-03-04
Beta-GH-hCoV-19/India|EPI ISL 2504562|2021-03-04
Beta-GH-hCoV-19/India|EPI ISL 2014132|2021-03
Beta-GH-hCoV-19/India|EPI ISL 2014131|2021-02
Beta-GH-hCoV-19/India|EPI ISL 2014135|2021-03
Beta-GH-hCoV-19/India|EPI ISL 2325060|2021
Beta-GH-hCoV-19/India|EPI ISL 2325057|2021-02
Beta-GH-hCoV-19/India|EPI ISL 2036295|2021-03
Beta-GH-hCoV-19/India|EPI ISL 2014133|2021
Beta-GH-hCoV-19/India|EPI ISL 2036297|2021
GK-hCoV-19/India|EPI ISL 1663516|2020-12-12
GK-hCoV-19/India|EPI ISL 1663555|2021-03-25
GK-hCoV-19/India|EPI ISL 1704626|2021-03-15
GK-hCoV-19/India|EPI ISL 1704627|2021-03-15
GK-hCoV-19/India|EPI ISL 1704625|2021-03-15
GK-hCoV-19/India|EPI ISL 1704561|2021-03-07
GK-hCoV-19/India|EPI ISL 1663564|2021-03-29
GK-hCoV-19/India|EPI ISL 1663563|2021-03-25
GK-hCoV-19/India|EPI ISL 1663553|2021-03-25
GK-hCoV-19/India|EPI ISL 1663557|2021-03-25
GK-hCoV-19/India|EPI ISL 1663502|2021-03-28
GK-hCoV-19/India|EPI ISL 1663562|2021-03-25
GK-hCoV-19/India|EPI ISL 1663529|2021-03-25
GK-hCoV-19/India|EPI ISL 1663558|2021-03-25
GK-hCoV-19/India|EPI ISL 1704234|2021-03-08
GK-hCoV-19/India|EPI ISL 1704620|2021-03-11
GK-hCoV-19/India|EPI ISL 1704617|2021-03-11
GK-hCoV-19/India|EPI ISL 1704618|2021-03-11
GK-hCoV-19/India|EPI ISL 2426161|2021-04-18
GK-hCoV-19/India|EPI ISL 2426171|2021-04-19
GK-hCoV-19/India|EPI ISL 2426160|2021-04-17
Delta-G-hCoV/India|EPI ISL 2710009|2021-04-29
Delta-G-hCoV/India|EPI ISL 2710008|2021-04-28
Delta-G-hCoV/India|EPI ISL 2342870|2021-04-02
Delta-G-hCoV/India|EPI ISL 2426182|2021-04-24
Delta-G-hCoV/India|EPI ISL 2342976|2021-04-14
Delta-G-hCoV/India|EPI ISL 2546289|2021-03-22
Delta-G-hCoV/Germany|EPI ISL 3044890|2021-07-15
GV-hCoV-19/India|EPI ISL 2897810|2021-04-24
GV-hCoV-19/India|EPI ISL 3060737|2021-06-29
GV-hCoV-19/India|EPI ISL 1533775|2021-02-21
GV-hCoV-19/India|EPI ISL 2774997|2021-06-04
GV-hCoV-19/India|EPI ISL 3008137|2021-05-05
GV-hCoV-19/India|EPI ISL 2774698|2021-06-08
GV-hCoV-19/India|EPI ISL 2774665|2021-04-03
GV-hCoV-19/India|EPI ISL 2774999|2021-06-04
GV-hCoV-19/India|EPI ISL 2156993|2021-04-17
GV-hCoV-19/India|EPI ISL 3008145|2021-05-13

Figure 3: The Neighbor-Joining phylogenetic tree based on 127 SARS-CoV-2 genomes using NAAKV method (k=4).

NAAKV and the traditional k-mer method are presented in Fig. 3 and Fig. 4, respectively. NAAKV gets right analysis in a shorter time. Additionally, we can see that NAAKV is superior to the traditional k-mer method in the Delta-G clade.

3.2. Bacteria

There are a lot of bacteria living on the earth. They are the major representatives of the prokaryotes. Due to the large genome size of bacteria, although researchers want to study them through phylogenetic trees, the process of

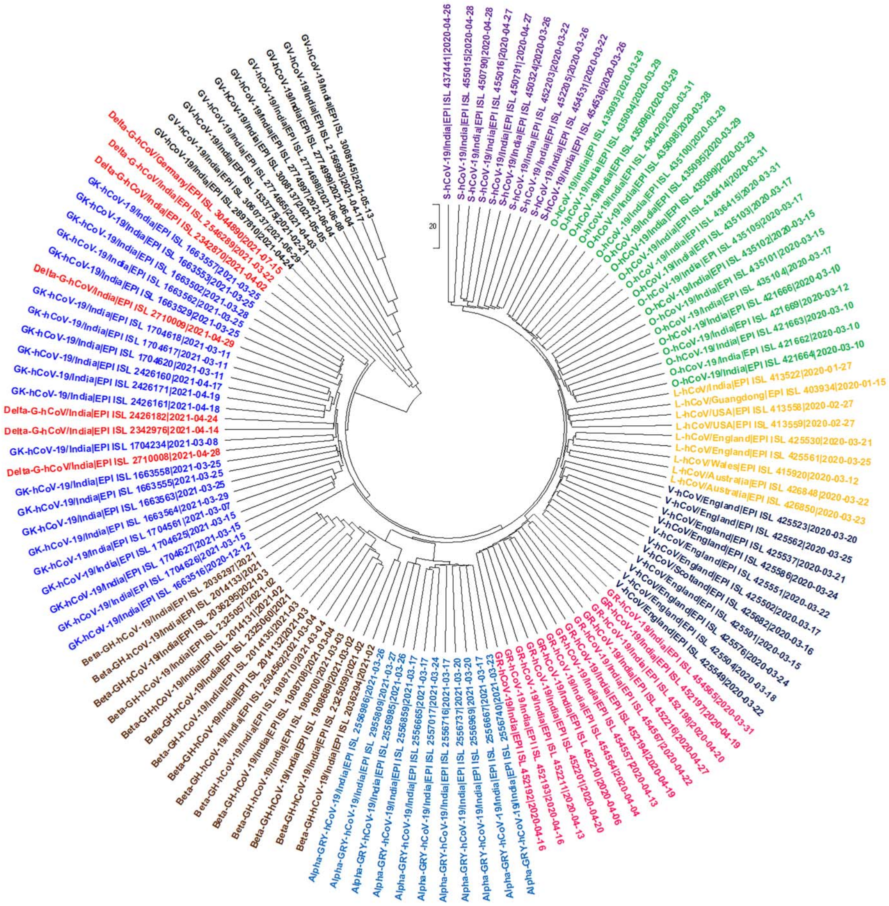
20
S-hCoV-19/India|EPI ISL 437441|2020-04-26
S-hCoV-19/India|EPI ISL 455015|2020-04-28
S-hCoV-19/India|EPI ISL 450790|2020-04-28
S-hCoV-19/India|EPI ISL 455016|2020-04-27
S-hCoV-19/India|EPI ISL 450791|2020-04-27
S-hCoV-19/India|EPI ISL 450324|2020-03-26
S-hCoV-19/India|EPI ISL 452203|2020-03-22
S-hCoV-19/India|EPI ISL 452205|2020-03-26
S-hCoV-19/India|EPI ISL 454531|2020-03-22
S-hCoV-19/India|EPI ISL 454536|2020-03-26
O-hCoV-19/India|EPI ISL 435093|2020-03-29
O-hCoV-19/India|EPI ISL 435094|2020-03-29
O-hCoV-19/India|EPI ISL 435096|2020-03-29
O-hCoV-19/India|EPI ISL 436420|2020-03-31
O-hCoV-19/India|EPI ISL 435098|2020-03-28
O-hCoV-19/India|EPI ISL 435100|2020-03-29
O-hCoV-19/India|EPI ISL 435095|2020-03-29
O-hCoV-19/India|EPI ISL 435099|2020-03-29
O-hCoV-19/India|EPI ISL 436414|2020-03-31
O-hCoV-19/India|EPI ISL 436415|2020-03-17
O-hCoV-19/India|EPI ISL 435103|2020-03-17
O-hCoV-19/India|EPI ISL 435105|2020-03-15
O-hCoV-19/India|EPI ISL 435102|2020-03-15
O-hCoV-19/India|EPI ISL 435101|2020-03-17
O-hCoV-19/India|EPI ISL 435104|2020-03-10
O-hCoV-19/India|EPI ISL 421666|2020-03-12
O-hCoV-19/India|EPI ISL 421669|2020-03-10
O-hCoV-19/India|EPI ISL 421663|2020-03-10
O-hCoV-19/India|EPI ISL 421662|2020-03-10
O-hCoV-19/India|EPI ISL 421664|2020-03-10
L-hCoV/India|EPI ISL 413522|2020-01-27
L-hCoV/Guangdong|EPI ISL 403934|2020-01-15
L-hCoV/USA|EPI ISL 413558|2020-02-27
L-hCoV/USA|EPI ISL 413559|2020-02-27
L-hCoV/England|EPI ISL 425530|2020-03-21
L-hCoV/England|EPI ISL 425561|2020-03-25
L-hCoV/Wales|EPI ISL 415920|2020-03-12
L-hCoV/Australia|EPI ISL 426848|2020-03-22
L-hCoV/Australia|EPI ISL 426850|2020-03-23
V-hCoV/England|EPI ISL 425523|2020-03-20
V-hCoV/England|EPI ISL 425562|2020-03-25
V-hCoV/England|EPI ISL 425537|2020-03-21
V-hCoV/England|EPI ISL 425586|2020-03-24
V-hCoV/England|EPI ISL 425551|2020-03-22
V-hCoV/Scotland|EPI ISL 425502|2020-03-17
V-hCoV/England|EPI ISL 425682|2020-03-16
V-hCoV/England|EPI ISL 425501|2020-03-15
V-hCoV/England|EPI ISL 425576|2020-03-24
V-hCoV/England|EPI ISL 425504|2020-03-18
V-hCoV/England|EPI ISL 425549|2020-03-22
GR-hCoV-19/India|EPI ISL 454565|2020-03-31
GR-hCoV-19/India|EPI ISL 452197|2020-04-19
GR-hCoV-19/India|EPI ISL 452198|2020-04-20
GR-hCoV-19/India|EPI ISL 454567|2020-04-27
GR-hCoV-19/India|EPI ISL 452216|2020-04-22
GR-hCoV-19/India|EPI ISL 452194|2020-04-19
GR-hCoV-19/India|EPI ISL 454557|2020-04-13
GR-hCoV-19/India|EPI ISL 454564|2020-04-04
GR-hCoV-19/India|EPI ISL 452210|2020-04-06
GR-hCoV-19/India|EPI ISL 452201|2020-04-20
GR-hCoV-19/India|EPI ISL 452211|2020-04-13
GR-hCoV-19/India|EPI ISL 452193|2020-04-16
GR-hCoV-19/India|EPI ISL 452192|2020-04-16
Alpha-GRY-hCoV-19/India|EPI ISL 2556740|2021-03-23
Alpha-GRY-hCoV-19/India|EPI ISL 2556667|2021-03-17
Alpha-GRY-hCoV-19/India|EPI ISL 2556969|2021-03-20
Alpha-GRY-hCoV-19/India|EPI ISL 2556737|2021-03-20
Alpha-GRY-hCoV-19/India|EPI ISL 2556716|2021-03-17
Alpha-GRY-hCoV-19/India|EPI ISL 2557017|2021-03-24
Alpha-GRY-hCoV-19/India|EPI ISL 2556665|2021-03-17
Alpha-GRY-hCoV-19/India|EPI ISL 2556859|2021-03-17
Alpha-GRY-hCoV-19/India|EPI ISL 2556985|2021-03-26
Alpha-GRY-hCoV-19/India|EPI ISL 2955809|2021-03-27
Alpha-GRY-hCoV-19/India|EPI ISL 2556986|2021-03-26
Beta-GH-hCoV-19/India|EPI ISL 2036294|2021-02
Beta-GH-hCoV-19/India|EPI ISL 2325059|2021-02
Beta-GH-hCoV-19/India|EPI ISL 1908689|2021-03-02
Beta-GH-hCoV-19/India|EPI ISL 1908700|2021-03-03
Beta-GH-hCoV-19/India|EPI ISL 1908708|2021-03-04
Beta-GH-hCoV-19/India|EPI ISL 1908710|2021-03-04
Beta-GH-hCoV-19/India|EPI ISL 2504562|2021-03-04
Beta-GH-hCoV-19/India|EPI ISL 2014132|2021-03
Beta-GH-hCoV-19/India|EPI ISL 2014135|2021-03
Beta-GH-hCoV-19/India|EPI ISL 2325060|2021
Beta-GH-hCoV-19/India|EPI ISL 2014131|2021-02
Beta-GH-hCoV-19/India|EPI ISL 2325057|2021-02
Beta-GH-hCoV-19/India|EPI ISL 2036295|2021-03
Beta-GH-hCoV-19/India|EPI ISL 2014133|2021
Beta-GH-hCoV-19/India|EPI ISL 2036297|2021
GK-hCoV-19/India|EPI ISL 1663516|2020-12-12
GK-hCoV-19/India|EPI ISL 1704626|2021-03-15
GK-hCoV-19/India|EPI ISL 1704627|2021-03-15
GK-hCoV-19/India|EPI ISL 1704625|2021-03-15
GK-hCoV-19/India|EPI ISL 1704561|2021-03-07
GK-hCoV-19/India|EPI ISL 1663564|2021-03-29
GK-hCoV-19/India|EPI ISL 1663563|2021-03-25
GK-hCoV-19/India|EPI ISL 1663555|2021-03-25
GK-hCoV-19/India|EPI ISL 1663558|2021-03-25
Delta-G-hCoV/India|EPI ISL 2710008|2021-04-28
GK-hCoV-19/India|EPI ISL 1704234|2021-03-08
Delta-G-hCoV/India|EPI ISL 2342976|2021-04-14
Delta-G-hCoV/India|EPI ISL 2426182|2021-04-24
GK-hCoV-19/India|EPI ISL 2426161|2021-04-18
GK-hCoV-19/India|EPI ISL 2426171|2021-04-19
GK-hCoV-19/India|EPI ISL 2426160|2021-04-17
GK-hCoV-19/India|EPI ISL 1704620|2021-03-11
GK-hCoV-19/India|EPI ISL 1704617|2021-03-11
GK-hCoV-19/India|EPI ISL 1704618|2021-03-11
Delta-G-hCoV/India|EPI ISL 2710009|2021-04-29
GK-hCoV-19/India|EPI ISL 1663529|2021-03-25
GK-hCoV-19/India|EPI ISL 1663562|2021-03-28
GK-hCoV-19/India|EPI ISL 1663502|2021-03-25
GK-hCoV-19/India|EPI ISL 1663553|2021-03-25
GK-hCoV-19/India|EPI ISL 1663557|2021-03-25
Delta-G-hCoV/India|EPI ISL 2546289|2021-04-02
Delta-G-hCoV/India|EPI ISL 2342870|2021-03-22
Delta-G-hCoV/Germany|EPI ISL 3044890|2021-07-15
GV-hCoV-19/India|EPI ISL 2897810|2021-04-24
GV-hCoV-19/India|EPI ISL 3060737|2021-06-29
GV-hCoV-19/India|EPI ISL 1533775|2021-02-21
GV-hCoV-19/India|EPI ISL 2774665|2021-04-03
GV-hCoV-19/India|EPI ISL 3008137|2021-05-05
GV-hCoV-19/India|EPI ISL 2774997|2021-06-04
GV-hCoV-19/India|EPI ISL 2774698|2021-06-08
GV-hCoV-19/India|EPI ISL 2774999|2021-06-04
GV-hCoV-19/India|EPI ISL 2156993|2021-04-17
GV-hCoV-19/India|EPI ISL 3008145|2021-05-13

Figure 4: The Neighbor-Joining phylogenetic tree based on 127 SARS-CoV-2 genomes using the traditional k-mer method (k=7).

establishing evolutionary tree is very difficult using traditional multiple sequence alignment methods with current computational algorithms. A bacterial dataset containing 59 strains with long genomes ranging from 0.8 to 5 million bp is using to do research. As shown in Fig. 5, with $k = 2$ the sequences are quickly and correctly divided into 14 separate families by our NAAKV method. Compared with the traditional k-mer(with $k = 9$) method Fig. 6, our method has obvious advantages in time, see Table 4.

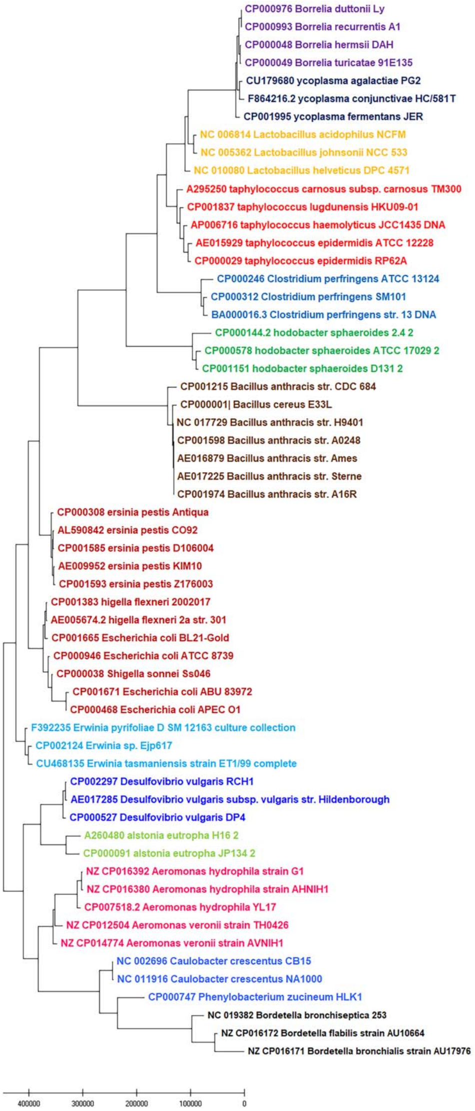
CP000976 Borrelia duttonii Ly
CP000993 Borrelia recurrentis A1
CP000048 Borrelia hermsii DAH
CP000049 Borrelia turicatae 91E135
CU179680 ycoplasma agalactiae PG2
F864216.2 ycoplasma conjunctivae HC/581T
CP001995 ycoplasma fermentans JER
NC 006814 Lactobacillus acidophilus NCFM
NC 005362 Lactobacillus johnsonii NCC 533
NC 010080 Lactobacillus helveticus DPC 4571
A295250 taphylococcus carnosus subsp. carnosus TM300
CP001837 taphylococcus lugdunensis HKU09-01
AP006716 taphylococcus haemolyticus JCC1435 DNA
AE015929 taphylococcus epidermidis ATCC 12228
CP000029 taphylococcus epidermidis RP62A
CP000246 Clostridium perfringens ATCC 13124
CP000312 Clostridium perfringens SM101
BA000016.3 Clostridium perfringens str. 13 DNA
CP000144.2 hodobacter sphaeroides 2.4 2
CP000578 hodobacter sphaeroides ATCC 17029 2
CP001151 hodobacter sphaeroides D131 2
CP001215 Bacillus anthracis str. CDC 684
CP000001| Bacillus cereus E33L
NC 017729 Bacillus anthracis str. H9401
CP001598 Bacillus anthracis str. A0248
AE016879 Bacillus anthracis str. Ames
AE017225 Bacillus anthracis str. Sterne
CP001974 Bacillus anthracis str. A16R
CP000308 ersinia pestis Antiqua
AL590842 ersinia pestis CO92
CP001585 ersinia pestis D106004
AE009952 ersinia pestis KIM10
CP001593 ersinia pestis Z176003
CP001383 higella flexneri 2002017
AE005674.2 higella flexneri 2a str. 301
CP001665 Escherichia coli BL21-Gold
CP000946 Escherichia coli ATCC 8739
CP000038 Shigella sonnei Ss046
CP001671 Escherichia coli ABU 83972
CP000468 Escherichia coli APEC O1
F392235 Erwinia pyrifoliae D SM 12163 culture collection
CP002124 Erwinia sp. Ejp617
CU468135 Erwinia tasmaniensis strain ET1/99 complete
CP002297 Desulfovibrio vulgaris RCH1
AE017285 Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough
CP000527 Desulfovibrio vulgaris DP4
A260480 alstonia eutropha H16 2
CP000091 alstonia eutropha JP134 2
NZ CP016392 Aeromonas hydrophila strain G1
NZ CP016380 Aeromonas hydrophila strain AHNIH1
CP007518.2 Aeromonas hydrophila YL17
NZ CP012504 Aeromonas veronii strain TH0426
NZ CP014774 Aeromonas veronii strain AVNIH1
NC 002696 Caulobacter crescentus CB15
NC 011916 Caulobacter crescentus NA1000
CP000747 Phenylobacterium zucineum HLK1
NC 019382 Bordetella bronchiseptica 253
NZ CP016172 Bordetella flabilis strain AU10664
NZ CP016171 Bordetella bronchialis strain AU17976
400000
300000
200000
100000
0

Figure 5: The Neighbor-Joining phylogenetic tree based on 59 bacteria genomes using NAAKV method (k=2).

3.3. Hepatitis C virus

Hepatitis C is a humoral infectious disease caused by hepatitis C virus (HCV), which mainly influences the liver. The initial treatment depends

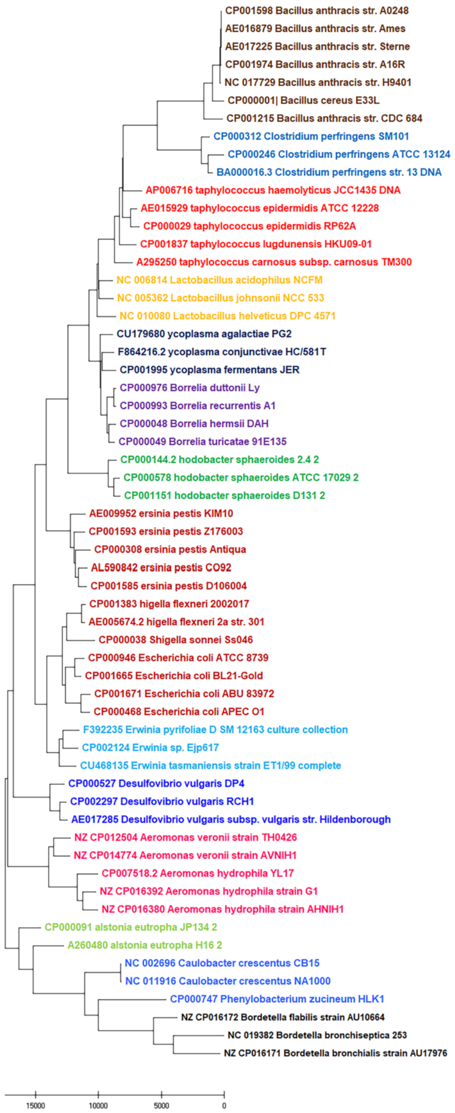
CP001598 Bacillus anthracis str. A0248
AE016879 Bacillus anthracis str. Ames
AE017225 Bacillus anthracis str. Sterne
CP001974 Bacillus anthracis str. A16R
NC 017729 Bacillus anthracis str. H9401
CP000001| Bacillus cereus E33L
CP001215 Bacillus anthracis str. CDC 684
CP000312 Clostridium perfringens SM101
CP000246 Clostridium perfringens ATCC 13124
BA000016.3 Clostridium perfringens str. 13 DNA
AP006716 taphylococcus haemolyticus JCC1435 DNA
AE015929 taphylococcus epidermidis ATCC 12228
CP000029 taphylococcus epidermidis RP62A
CP001837 taphylococcus lugdunensis HKU09-01
A295250 taphylococcus carnosus subsp. carnosus TM300
NC 006814 Lactobacillus acidophilus NCFM
NC 005362 Lactobacillus johnsonii NCC 533
NC 010080 Lactobacillus helveticus DPC 4571
CU179680 ycoplasma agalactiae PG2
F864216.2 ycoplasma conjunctivae HC/581T
CP001995 ycoplasma fermentans JER
CP000976 Borrelia duttonii Ly
CP000993 Borrelia recurrentis A1
CP000048 Borrelia hermsii DAH
CP000049 Borrelia turicatae 91E135
CP000144.2 hodobacter sphaeroides 2.4 2
CP000578 hodobacter sphaeroides ATCC 17029 2
CP001151 hodobacter sphaeroides D131 2
AE009952 ersinia pestis KIM10
CP001593 ersinia pestis Z176003
CP000308 ersinia pestis Antiqua
AL590842 ersinia pestis CO92
CP001585 ersinia pestis D106004
CP001383 higella flexneri 2002017
AE005674.2 higella flexneri 2a str. 301
CP000038 Shigella sonnei Ss046
CP000946 Escherichia coli ATCC 8739
CP001665 Escherichia coli BL21-Gold
CP001671 Escherichia coli ABU 83972
CP000468 Escherichia coli APEC O1
F392235 Erwinia pyrifoliae D SM 12163 culture collection
CP002124 Erwinia sp. Ejp617
CU468135 Erwinia tasmaniensis strain ET1/99 complete
CP000527 Desulfovibrio vulgaris DP4
CP002297 Desulfovibrio vulgaris RCH1
AE017285 Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough
NZ CP012504 Aeromonas veronii strain TH0426
NZ CP014774 Aeromonas veronii strain AVNIH1
CP007518.2 Aeromonas hydrophila YL17
NZ CP016392 Aeromonas hydrophila strain G1
NZ CP016380 Aeromonas hydrophila strain AHNIH1
CP000091 alstonia eutropha JP134 2
A260480 alstonia eutropha H16 2
NC 002696 Caulobacter crescentus CB15
NC 011916 Caulobacter crescentus NA1000
CP000747 Phenylobacterium zucineum HLK1
NZ CP016172 Bordetella flabilis strain AU10664
NC 019382 Bordetella bronchiseptica 253
NZ CP016171 Bordetella bronchialis strain AU17976
15000
10000
5000
0

Figure 6: The Neighbor-Joining phylogenetic tree based on 59 bacteria genomes using the traditional k-mer method (k=9).

on the type of HCV. Thus, the identification of the type of HCV present is critically important. Using the NAAKV method, we divide 82 gene components into 6 genotypes at $k = 4$ as shown in Fig. 7. In contrast, as can be seen in Fig. 8, some genomes of type 6 are not perfectly sorted by the

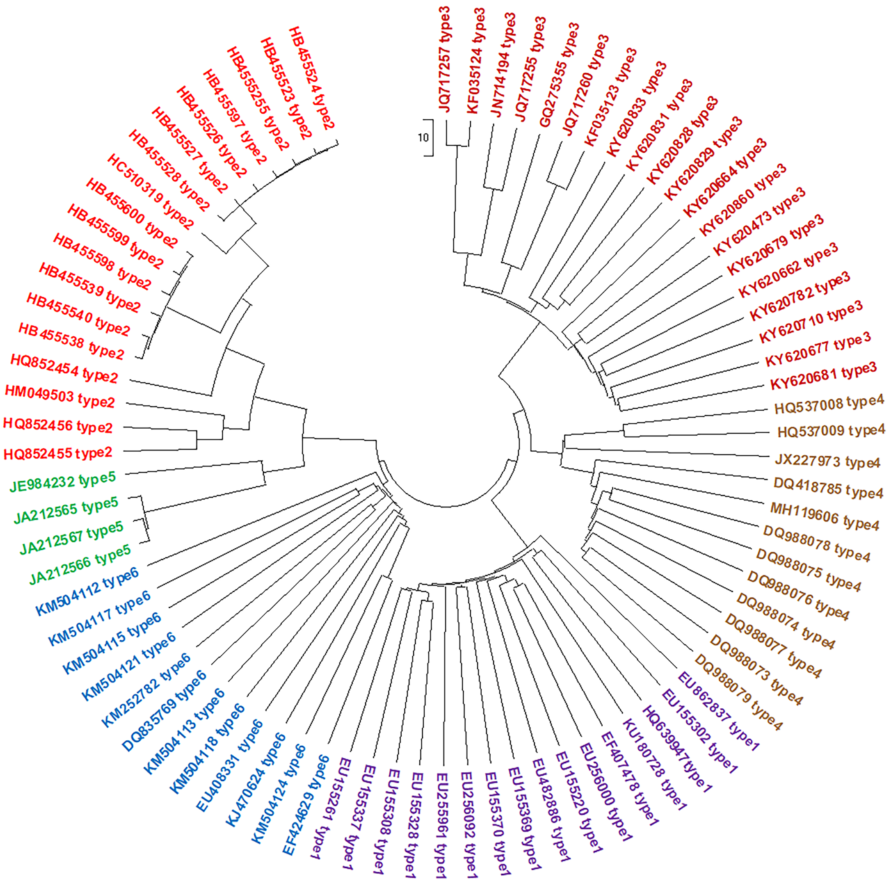

Figure 7: The Neighbor-Joining phylogenetic tree based on 82 HCV genomes using NAAKV method (k=4).

traditional k-mer method, such as EU408331, KM504118 and KM504112. Moreover, our method takes far less time. We also analyze it with MUSCLE method as shown in supplementary Fig. S1.

3.4. Human Rhinovirus

Human rhinovirus (HRV) is the main pathogen responsible for the common cold in humans. Previous works divided HRV into three categories: HRV-A, HRV-B, and HRV-C [24, 25]. We assemble a dataset consulting of 113 HRVs and 3 HEV-C complete genes [26]. As shown in Fig. 9, NAAKV successfully classify 116 HRVs into four categories with k=2. MUSCLE also gets the right

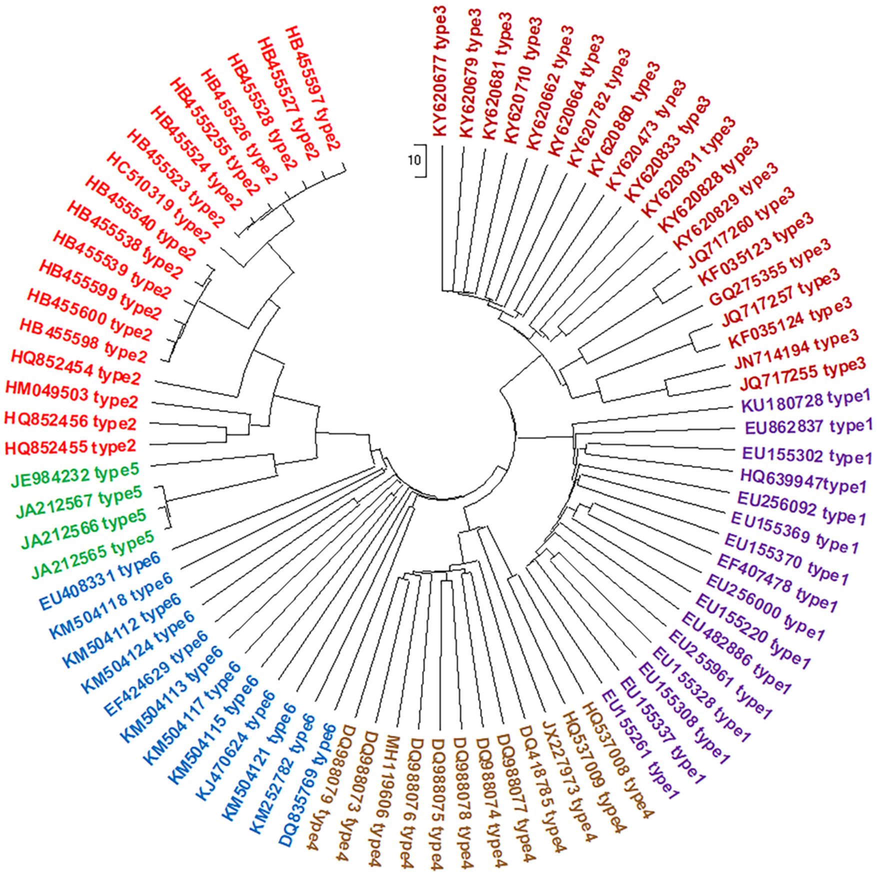

Figure 8: The Neighbor-Joining phylogenetic tree based on 82 HCV genomes using the traditional k-mer method (k=6).

result as shown in Fig. 10. It is worth noting that NAAKV is faster than MUSCLE. The result of traditional k-mer method is shown in supplementary Fig. S2.

3.5. Hepatitis B virus

Hepatitis B virus is the pathogen causing hepatitis B. Hepatitis B is a viral infection that attacks the liver and can cause both acute and chronic disease [27]. As our method can covert the DNA sequence to a PAAS, it can reflect the strong correlation between the amino acids. To demonstrate this advantage, 151 complete HBV genomes with our NAAKV method and

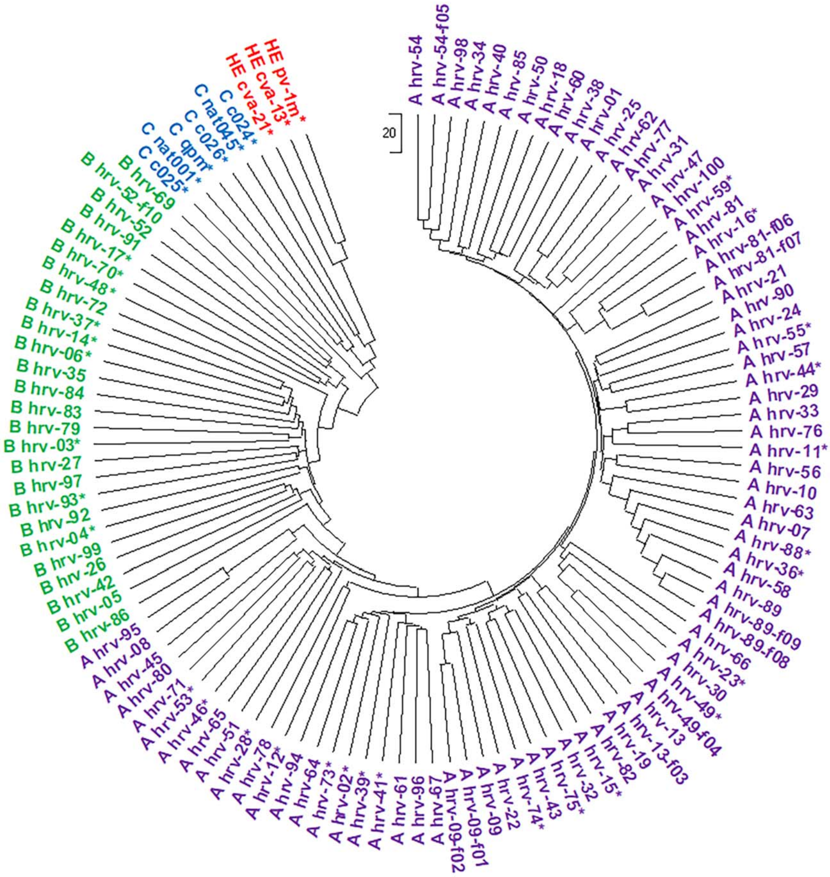

Figure 9: The Neighbor-Joining phylogenetic tree based on 114 HRV genomes using NAAKV method (k=2).

their corresponding protein sequence with k-mer method are employed to do the research. The result based on NAAKV can be seen in the Fig. 11, which shows the elegant phylogenetic tree. However, many classes are mixed together in Fig. 12, which is built by k-mer method and protein sequences.

4. Discussion and conclusion

This study sets out with the aim of raising a novel genome classification method called NAAKV. Results of this study confirm that NAAKV can achieve accurate evolutionary analysis and subtype classification on SARS-CoV-2, bacteria, HCV, HRV and HBV. Besides, NAAKV saves more time

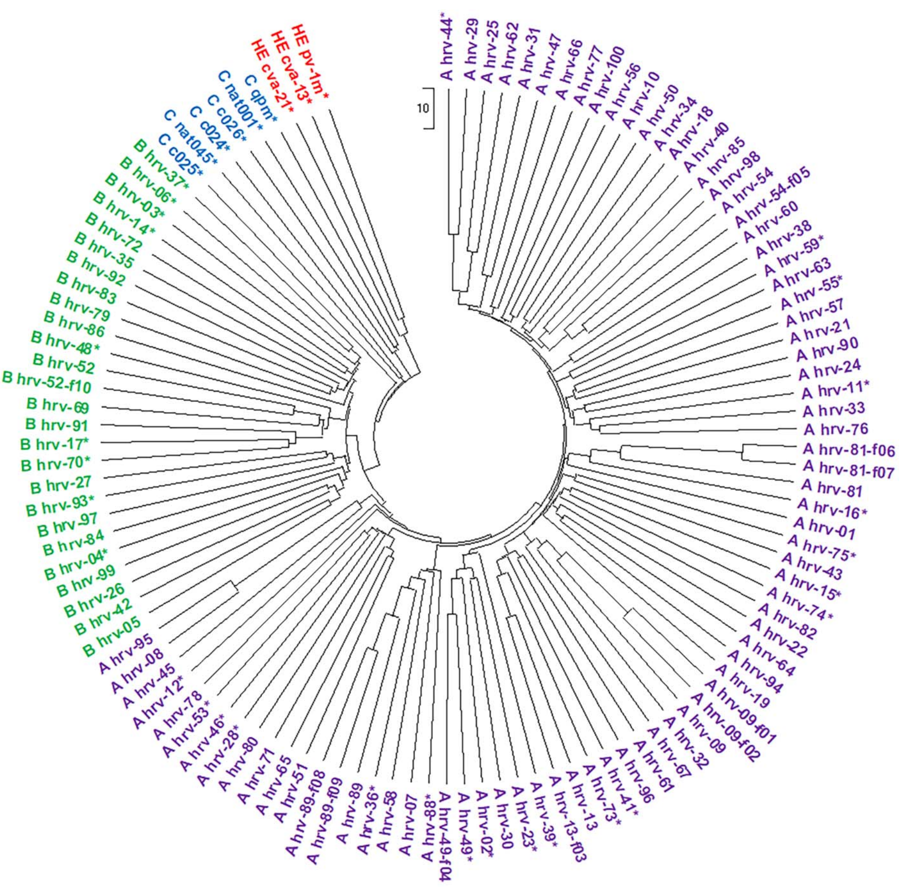

Figure 10: The Neighbor-Joining phylogenetic tree based on 114 HRV genomes using MUSCLE.

than k-mer and MUSCLE method when the same correct results are acquired.

It is different from the protein sequence. In our algorithm, one step is to convert the DNA sequence into a PAAS by moving one nucleotide each time, and we do not need to find the coding region of gene. Consequently, we can avoid the problem that some sequences do not have protein sequences. In the HBV dataset, we compare protein sequence with our method, and show the corresponding superiority. Meanwhile, to validate the advantage for moving one nucleotide each time, we compute the evolutionary tree based on pseudo amino acid sequences by moving two or three cases in the HCV dataset, and the results are shown in supplementary Fig. S3 and Fig. S4 respectively.

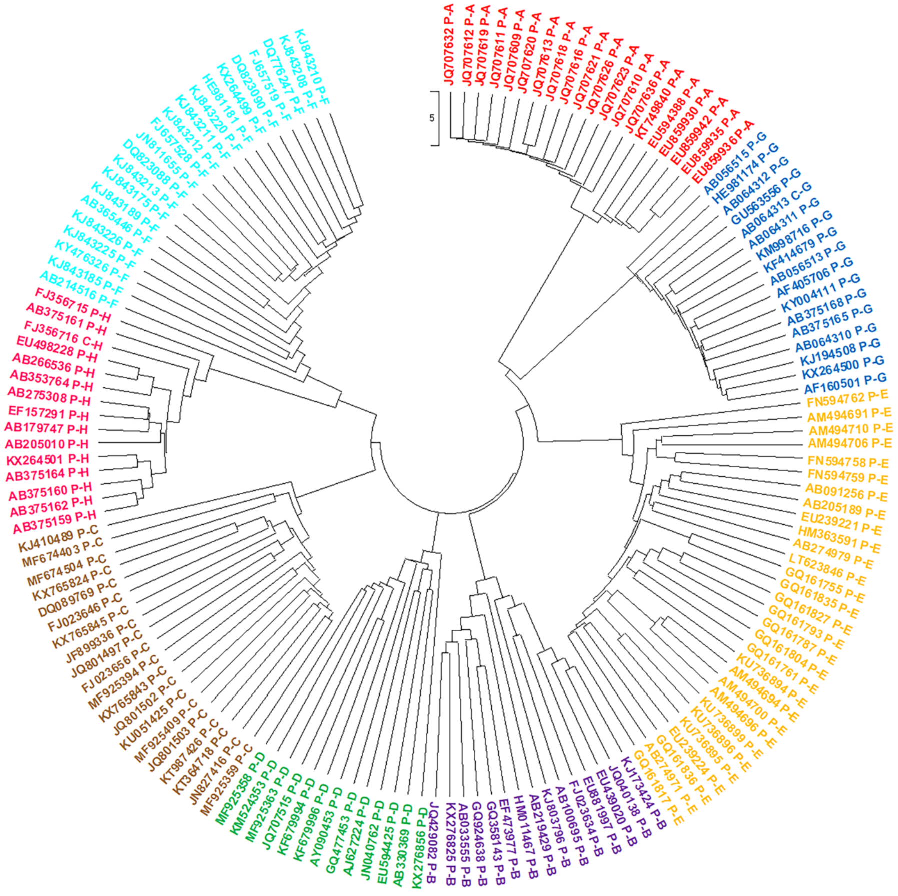

Figure 11: The Neighbor-Joining phylogenetic tree based on 151 HBV genomes using NAAKV method (k=2).

To validate the advantage using PAAS, we evaluate the k-mer types based on PAAS in the bacteria dataset mentioned in the Result section and compare the result based on protein sequence. The number of k-mer types is shown in Table 5. We can see that the k-mer types occurring in the bacteria dataset based on PAAS are lower much than that based on PS. Apparently, given a k value, the number of k-mer types of PAAS is far less than protein sequence.

In summary, NAAKV is not only a very fast but also an accurate method of genome comparison.

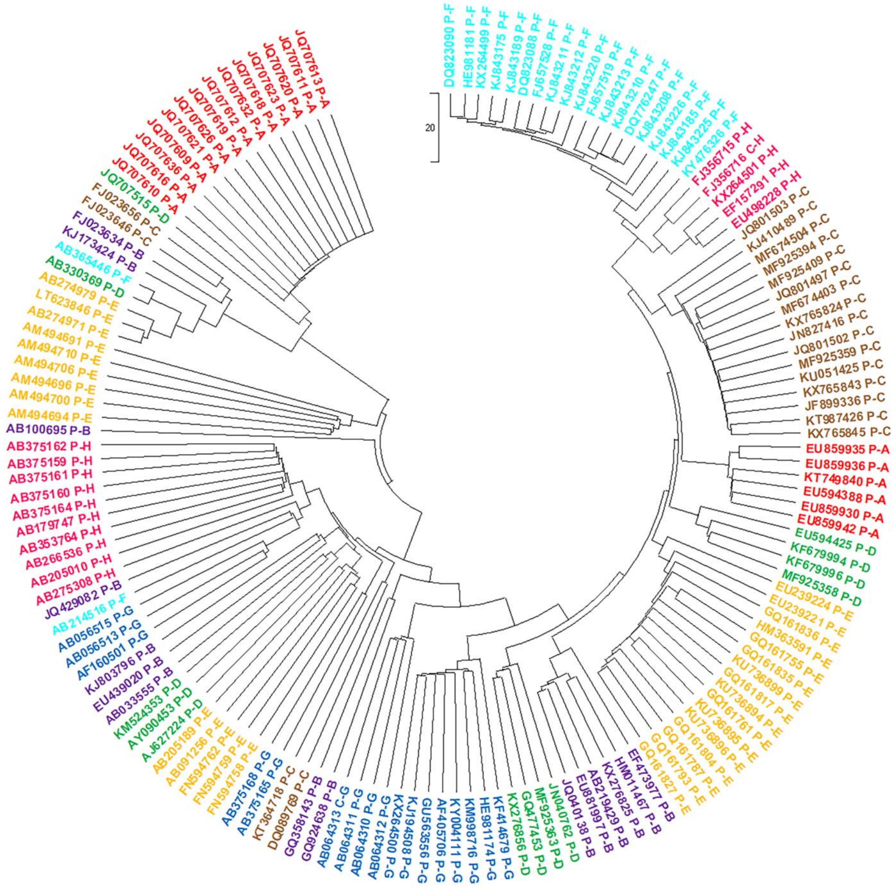

Figure 12: The Neighbor-Joining phylogenetic tree based on 151 HBV protein sequences using the traditional k-mer method (k=2).

Table 5: The actual number of k-mer types in a bacteria dataset based on pseudo amino acid sequence (PAAS) and protein sequence (PS) respectively

k	2	3	4	5	6	7	8
PAAS	95	448	2,015	9,104	41,059	177,658	637,319
PS	400	8,000	146,576	907,888	1,548,488	1,709,253	1,735,029

Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC) grant 12171275, 11871093, Tsinghua University Spring Breeze Fund

(2020Z99CFY044), Tsinghua University start-up fund, Tsinghua University Education Foundation fund (042202008), Promotion plan for young teachers' scientific research ability of Beijing University of Civil Engineering and Architecture (X21026), R&D Program of Beijing Municipal Education Commission (KM202210016014). Professor Stephen Shing-Toung Yau is grateful to the National Center for Theoretical Sciences (NCTS) for providing an excellent research environment while part of this research was done.

References

[1] H. Lin, W. Liu, H. Gao, J. Nie and Q. Fan, Trends in Transmissibility of 2019 Novel Coronavirus-Infected Pneumonia in Wuhan and 29 Provinces in China. *SSRN Electronic Journal.* MR4128866

[2] P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang, H.-R. Si, Y. Zhu, B. Li, C.-L. Huang, H.-D. Chen, J. Chen, Y. Luo, H. Guo, R.-D. Jiang, M.-Q. Liu, Y. Chen, X.-R. Shen, X. Wang, X.-S. Zheng, K. Zhao, Q.-J. Chen, F. Deng, L.-L. Liu, B. Yan, F.-X. Zhan, Y.-Y. Wang, G.-F. Xiao and Z.-L. Shi, A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579** (2020), 270–273. MR2265238

[3] E. B. Hodcroft, M. Zuber, S. Nadeau, T. G. Vaughan and R. A. Neher, Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* **595** (2021), 707–712.

[4] M. Makoni, South Africa responds to new SARS-CoV-2 variant. *The Lancet* **397** (2021), 267.

[5] C. Santos, N. Muoz, S. Klug, M. Almonte, I. Guerrero, M. Alvarez, C. Velarde, O. Galdos, M. Castillo and J. Walboomers, HPV types and cofactors causing cervical cancer in Peru. *British Journal of Cancer* **85** (2001), 966–971.

[6] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, Basic local alignment search tool. *Journal of Molecular Biology* **215(3)** (1990), 403–410.

[7] J. D. Thompson, D. G. Higgins and T. J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22** (1994), 1673–1680.

[8] R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32** (2004), 1792–7.

[9] B. Guillaume, C. C. Xin, C. Yao-Ban, C. Xin-Yi, Y. Cong, J. M. Hogan, S. R. Maetschke and M. A. Ragan, Alignment-free inference of hierarchical and reticulate phylogenomic relationships. *Briefings in Bioinformatics* **20** (2019), 426–435.

[10] I. Ulitsky, D. Burstein, T. Tuller and B. Chor, The Average Common Substring Approach to Phylogenomic Reconstruction. *Journal of Computational Biology* **13(2)** (2006), 336–350. MR2255263

[11] H. Bernhard, K. Fabian and P. Peter. andi: Fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics* **31(8)** (2015), 1169–1175.

[12] L. He, Y. Li, R. L. He and S. S.-T. Yau, A novel alignment-free vector method to cluster protein sequences. *Journal of Theoretical Biology* **427** (2017), 41–52. MR3665144

[13] K. D. Murray, W. Christfried, O. C. Soon, B. Justin, W. Norman and P. Andreas, kWIP: The k-mer weighted inner product, a de novo estimator of genetic similarity. *PLoS Computational Biology* **13(9)** (2017), e1005727.

[14] L. He, R. Dong, R. L. He and S. S.-T. Yau, Positional Correlation Natural Vector: A Novel Method for Genome Comparison. *International Journal of Molecular Sciences* **21(11)** (2020), 3859.

[15] B. E. Blaisdell, 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. *Journal of Molecular Evolution* **29** (1989), 538–547.

[16] G. E. Sims, S. R. Jun, G. A. Wu and S. H. Kim, Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences of the United States of America* **106** (2009), 2677–2682.

[17] J. J. Choi and S. H. Kim, A genome Tree of Life for the Fungi kingdom. *Proceedings of the National Academy of Sciences* **114** (2017), 9391–9396.

[18] M. Guillaume, C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27(6)** (2011), 764–770.

[19] G.-B. Han and D.-H. Cho, Genome classification improvements based on k-mer intervals in sequences. *Genomics* **111** (2019), 1574–1582.

[20] S. V. Thankachan, S. P. Chockalingam and Y. Liu, A greedy alignment-free distance estimator for phylogenetic inference. *BMC Bioinformatics* **18(8 Supplement)** (2017), 238.

[21] P. Kolekar, M. Kale and Kulkarni-Kale Urmila, Alignment-free distance measure based on return time distribution for sequence analysis: Applications to clustering, molecular phylogeny and subtyping. *Molecular Phylogenetics and Evolution* **65(2)** (2012), 510–522.

[22] A. Zielezinski, H. Z. Girgis, G. Bernard, C. A. Leimeister and W. M. Karlowski, Benchmarking of alignment-free sequence comparison methods. *Genome Biology* **20** (2019), 114.

[23] K. Sudhir, G. Stecher, M. Li, C. Knyaz and K. Tamura, MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* **35** (2018), 6.

[24] M. Deng, C. Yu, Q. Liang, R. L. He and S. S.-T. Yau, A Novel Method of Characterizing Genetic Sequences: Genome Space with Biological Distance and Applications. *PloS one* **6(3)** (2011), e17293.

[25] T. Hoang, C. Yin, H. Zheng, C. Yu, R. L. He and S. S.-T. Yau, A new method to cluster DNA sequences using Fourier power spectrum. *Journal of Theoretical Biology* **372** (2015), 135–145. MR3331829

[26] A. C. Palmenberg, D. Spiro, R. Kuzmickas, S. Wang, A. Djikeng, J. A. Rathe, C. M. Fraser-Liggett and S. B. Liggett, Sequencing and Analyses of All Known Human Rhinovirus Genomes Reveal Structure and Evolution. *Science* **324** (2009), 55–59.

[27] J. V. Lazarus, B. Timothy, B. Christian, K. Anna, M. Veronica, N. Michael, P. Capucine, P. Ulrike, R. Homie and L. A.Thomas, The hepatitis B epidemic and the urgent need for cure preparedness. *Nature Reviews Gastroenterology and Hepatology* **15** (2018), 517–518.

Xiaona Bao
School of Science
Beijing University of Civil Engineering and Architecture
Beijing 102616 China
E-mail address: 2102520020002@stu.bucea.edu.cn

Lily He
School of Science
Beijing University of Civil Engineering and Architecture
Beijing 102616 China
E-mail address: helili@bucea.edu.cn

Jingan Cui
School of Science
Beijing University of Civil Engineering and Architecture
Beijing 102616 China
E-mail address: cuijingan@bucea.edu.cn

Stephen S.-T. Yau
Department of Mathematical Sciences
Tsinghua University
Beijing 100084 China
E-mail address: yau@uic.edu

Received December 22, 2021