Communications in Information and Systems
Volume 19, Number 1, 17–35, 2019

Assessment of kmer degeneration method for complicated genomes

Shuai Liu†, Shaojun Pei†, Stephen S.-T. Yau, and Qi Wu

The kmer frequency is widely used in alignment-free sequence analysis methods. To better describe the overall statistical features of a complicated sequence, such as those of mammals, a longer length of kmer is required. However, the long length of kmer will cause exponential increasing of the types of kmer (4^K types of kmer with length K), which results in an extremely intensive computational burden and makes k-mer method impractical. In this work, we propose a novel method of kmer degeneration (KD) to balance the kmer length and kmer type. The method only considers N positions of nucleotides out of K positions of K-mers and degenerates all other $(K-N)$ positions. Then the K-mers can be substituted by $(K)N$-mer. Therefore, the kmer types were reduced from 4^K to 4^N and remain the linkages among the nucleotides within the K-mer. We first show how N can be determined for a given K. Then we assess which types of combinations of the N positions from the K positions are better for describing the sequence. Finally, to illustrate the utility of the method, we construct the phylogenetics tree of *Carnivora* with 16 genomes using our method, which is better than non-degenerated kmer with the same N value.

1. Introduction

Alignment-free (AF) sequence analysis methods focus on subsequences (words or kmers) of defined or varied lengths of the sequence [3, 4, 14]. The types[2], frequencies [13], or positions [5] of the kmers in the sequence are considered as characteristics to do genome comparison. Of the possible characteristics, frequency is the most widely used in alignment-free methods [4, 12, 13]. The algorithm was shown with the CVTree program as an example [13]. Starting with a given DNA sequence of length L, a sliding window

†Shaojun Pei and Shuai Liu contributed equally.
This paper was revised on May 10, 2019.

of length K is run through the position 1 to $L-K+1$ to count the frequency of subsequences (kmers or strings). The total possible types of such strings could be 4^K for DNA sequences. Denote the frequency of appearance of kmer $\alpha_1\alpha_2\cdots\alpha_K$ by $f(\alpha_1\alpha_2\cdots\alpha_K)$, where $\alpha_i \in \{A, C, G, T\}$. This frequency divided by the total number $L-K+1$ ($K << L$) of K-mers in the given DNA sequence may be taken as the probability $p(\alpha_1\alpha_2\cdots\alpha_K)$ of appearance of the kmer $\alpha_1\alpha_2\cdots\alpha_K$ in the sequence:

$$p(\alpha_1\alpha_2\cdots\alpha_K) = \frac{f(\alpha_1\alpha_2\cdots\alpha_K)}{K-L+1} \tag{1}$$

The collection of such frequencies or probabilities describes the overall statistical features of the sequence concerned, which could then be used for further analysis such as whole-genome phylogeny reconstruction.

Clearly, in this approach, the most important parameter is the length of kmer. Therefore, the length of kmer, or K value, should be determined by some types of prior information to obtain optimal target sequence resolution. One straightforward limit is that the K value must be far less than the sequence length, or L value [13]. Sims et al. [12] derived the upper limits of kmer length by cumulative relative entropy (CRE) for a given symbol sequence, which represented the accuracy of predicting kmer frequencies for all lengths of kmer from Equation 1. They found that the length reaches the upper limit for use in genome comparison when the CRE approaches zero. In a qualitative view, the kmer length is related to both the sequence length and the kmer type number. With the long k-mer lengths, the kmer type number increases exponentially, which causes two problems. One problem is that a large number of kmer types will exceed the storage capacity of the computer. Another is that, if the number of kmer types far exceeds the amount of kmers appearing in the sequence, the majority of kmers will appear only once in non-repetitive sequences. This means, for any non-repetitive sequence, long kmer lengths result in a uniform distribution of the kmer frequencies.

In this work, we proposed k-mer degeneration (KD) method to balance the longer kmer length and the increased kmer types. In k-mer degeneration method, we only keep N ($N << K$) positions in the K positions of a sliding window, so that a K-mer is degenerated to an N-mer. Therefore, the kmer types are reduced from 4^K to 4^N. At the same time, the N positions of the $(K)N$-mer could have a span ranging of K. Therefore, the linkages among nucleotides within the K-mer is, to some extent, reflected through the degenerated $(K)N$-mer, so the information in the $(K)N$-mer could be greater than that in a naive N-mer. Using Shannon information [11] as a measurement, we assessed our method on the data set of mammalian

genomes. First, for a given K, we used distribution of Shannon information to determine which N is suitable for $(K)N$-mer. Then we accessed different types of combinations of the N positions from the K positions and the results indicated that the cases of continuous N positions possess significantly less information than the cases of dispersed N positions for a fixed K. Finally, we constructed the phylogenetics tree of *Carnivora* with 16 genomes by using our method in the form of dispersed $N(K)$-mer, which is better than those using non-degenerated kmer with the same N value.

2. Materials and Methods

2.1. Genome data collection

Genome sequences were downloaded from public database. The selected species included human (*Homo sapiens*) and mouse (*Mus muscullus*) from the superorder of *Euarchontoglires*, dog (*Canis lupus familiaris*) from the superorder of *Laurasiatheria*, elephant (*Loxodonta africana*) from the superorder of *Afrotheria*, and wallaby (*Macropus eugenii*) from the order of *Marsupialia*. Except for the wallaby, which is located at the basal branch of the mammal phylogenetic tree, the other 4 species belong to the placentals, which is at the crown of the mammalian tree. The sequence length was fixed to 50 Kbps in this work. For different genomes, 10 segments of genome sequence were independently sampled from the genome.

To construct phylogenetic tree and evaluate the performance, we picked the mouse and human genomes to make up an outgroup, and red panda (*Ailurus fulgen*), ferret(*Mustela putorius furo*), giant panda (*Ailuropoda melanoleuca*), tiger (*Panthera tigris*), polar bear (*Ursus maritimus*), cat (*Felis catus*), pacific walrus (*Odobenus rosmarus*), sea otter (*Enhydra lutris*), brown bear (*Ursus arctos*), weddell seal (*Leptonychotes weddellii*), cheetah (*Acinonyx jubatus*), puma (*Puma concolor*), leopard (*Panthera pardus*) from *Carnivora* (Table S3[1]).

2.2. kmer degeneration (KD) Method

For a DNA sequence with length L, a window with length K is run through the sequence from the beginning with steps of one nucleotide. One can obtain $L - K + 1$ kmers appearing in the sequence, which belong to T_K types of

[1]Tables S1–S4 are available at: `https://intlpress.com/site/pub/files/_supp/cis/2019/v19n1/CIS-2019-v19n1-s1.zip`

kmers. It is always the case that T_K is smaller than $L - K + 1$, since there must be some kmers appearing more than once in the sequence. At the same time, the total number of kmer types would be 4^K , giving

$$\begin{cases} T_K \leq L - K + 1 \\ T_K \leq 4^K \end{cases} \tag{2}$$

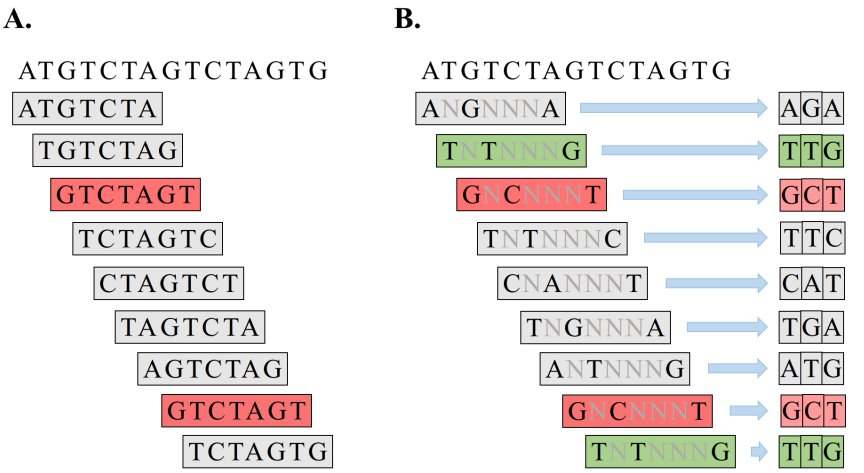


Figure 1: The kmer degeneration (KD) approach. The analysis of $K = 7$ kmer for a 15 bp sequence of "ATGTCTAGTCTAGTC" was introduced as a case. **A**. The standard kmer analysis approach. The two red boxes show that the two kmers with different starting points in the sequence belong to the same kmer type. **B**. The degeneration of these kmers. Three positions with the combination of 1-3-7 positions were used to degenerate every kmer. For each kmer, only the 3 positions were maintained, with the other 4 position degenerated as "N". The results of the degenerated $K(N)$-mer (in this case the 3-7-mer) are shown on the right. With the KD, one more 3-7-mer became identical, shown in green. The two different original 7-mers were "TGTCTAG" and "TCTAGTG". With KD they degenerated as the same 3-7-mer of "TTG". Thus, the kmer type was deduced.

In k-mer degeneration method, we only keep the N positions of nucleotide in K positions of the sliding window. Except for the N positions in the window, all other positions in the window are degenerated. A new $(K)N$-mer was then obtained, where the term "$(K)N$-mer" denotes the new kmer with length N that is degenerated other $(N - K)$ positions from the original K-mer (Figure 1). From a mathematical viewpoint, there are

$\binom{K}{N}$ combinations to pick N positions out of K positions. When dealing with one sequence, one can fix one combination and perform kmer degeneration (KD) method for a sequence. The symbol T_N is used to denote the number of $(K)N$-mer types in this case. Notably, although the number of kmers appearing in the sequence is the same for both K-mer and $(K)N$-mer, the total number of kmer types is decreased from 4^K to 4^N. Therefore, one has

$$\begin{cases} T_K \leq L - K + 1 \\ T_K \leq 4^K \\ T_N \leq 4^N \\ T_N \leq T_K \end{cases} \tag{3}$$

For a genome sequence segment with a length of 50 Kbp, one could analyze it with a kmer whose length is 50 bp. Consider the case that the 50-mer is degenerated into an 8 bp length of (50)8-mer, one has

$$\begin{cases} T_K \leq L - K + 1 = 49951 \\ T_K \leq 4^K = 4^{50} \approx 1.28 \times 10^{30} \\ T_N \leq 4^N = 4^8 = 65536 \\ T_N \leq T_K \end{cases} \tag{4}$$

Therefore, one can see the difference between using K-mers directly and instead using degenerated $(K)N$-mers to analyze sequence. The values of 49951 and 65536 are the same order of magnitude, but 1.28×10^{30} is much larger. Supposed 49951 balls were placed into 1.28×10^{30} boxes, each ball has an equal probability to be placed into any of the boxes. In most cases, one may obtain a distribution much closer to a uniform distribution, since it would be very rare to place two or more balls into one box. By contrast, if one put 49951 balls into 65536 boxes, one could easily obtain a recognizable distribution that would be significantly different from a uniform distribution(Figure 2).

2.3. Definition of the dispersed and continuous $(K)N$-mers

In kmer method, for a DNA sequence of length L, there is a sliding window of length K which is run through the sequence. Our kmer degeneration method only picks up N positions out of the K positions of the sliding window. To remain the span of K-mer, the first and last position of the window should be picked up. Then one can pick up other $N - 2$ positions randomly from remaining positions in the window. So there are $\binom{K-2}{N-2}$ combinations totally. In this study, we mainly studied two special cases. Two forms of

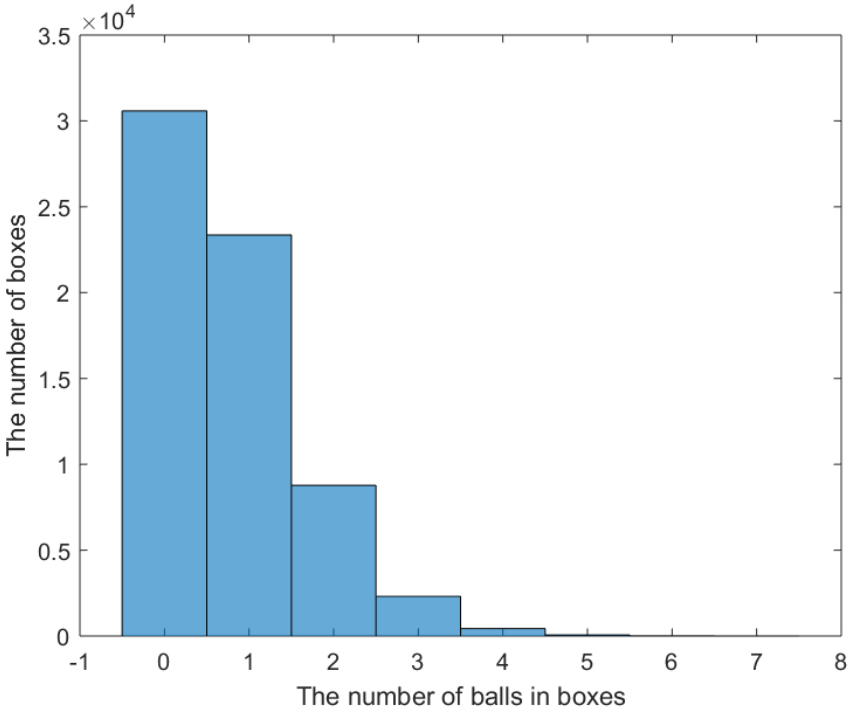


Figure 2: The distribution for 49951 balls into 65536. The horizontal axis is the number of balls in boxes. The vertical axis is the number of corresponding boxes.

$(K)N$-mers were defined — the dispersed form and continuous form. For any form of $(K)N$-mer, let D_i be the length of the interval between every two nucleotides, where $i = 1, 2, \ldots, N-1$. Then Let

$$d = \left\lfloor \frac{K-N}{N-1} \right\rfloor \tag{5}$$

Definition 1. If all the $d - 1 < D_i < d + 1$, $i = 1, 2, \ldots, N-1$, we call these $(K)N$-mers as the dispersed form.

For example, when $K = 7$ and $N = 3$, the d is $\frac{7-3}{3-1} = 2$; therefore, the three positions chosen for the $(K)N$-mer should be 1-3-7, 1-4-7, 1-5-7 (Fig. 3A). When $K = 8$ and $N = 3$, the d is also 2, the forms were: 1-3-8, 1-4-8, 1-5-8. With this method, we could produce a number of $(K)N$-mers as a group whose position was dispersed from the original K-mer.

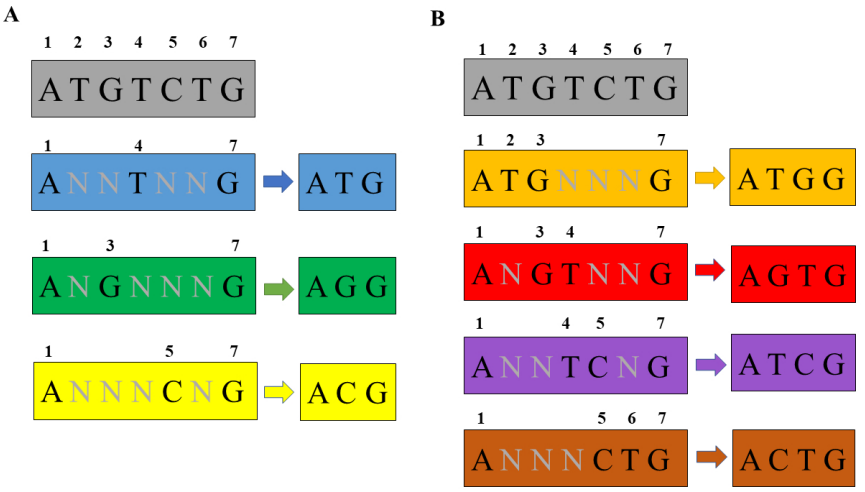


Figure 3: The dispersed form versus the continuous form of $K(N)$-mer. **A**. The dispersed form in the case of $K = 7$ and $N = 3$. **B**. The continuous form in the case of $K = 7$ and $N = 4$.

Definition 2. If all the centering $N - 2$ positions are continuous, we call these $(K)N$-mers as the continuous form.

For example, when $K = 7$ and $N = 4$, the continuous forms are 1-2-3-7, 1-3-4-7, 1-4-5-7, 1-5-6-7 (Fig. 3B).

2.4. Measurement for degenerated kmer method

As has been mentioned above, there are $\binom{K}{N}$ combinations for choosing N positions out of K positions. The question is whether there are any differences between these combination types. This question could be asked in the view of information theory: Which type of combination(s) could provide more information? The Shannon information is a well-defined index in this case. For a finite discrete random variable X with distribution $f(X) = p(x_i)$, where $i = 1, 2, \ldots, n$, the Shannon information $I(X)$ is

$$I(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{6}$$

When given a target sequence and the $(N)K$-mer, the number of $(N)K$-mer types T_N is determined. And the distribution is determined by all frequencies of the T_N types of kmer, so

$$I(N) = -\sum_{i=1}^{T_N} p_i \log_2 p_i \tag{7}$$

2.5. Generation of posterior distribution

The most straightforward measurement method would be to compute all of the combinations and compare them. However, for large values of K and N, such as the previous case of $K = 50$ and $N = 8$, the number of $\binom{K}{N}$ would be too large. We randomly sampled 1000 combinations from all combinations. For each type of combination of $(K)N$-mer, the frequencies of $(K)N$-mer were computed and then the Shannon information of this combination was calculated by the frequencies. The range from the maximum information value to the minimum information value was divided into 100 groups. The 1000 information values were then grouped to produce a posterior distribution of the information. With this posterior distribution, the differences among the combinations can be assessed.

2.6. Construction of phylogenetic trees

Phylogenetic tree is required to be constructed to test and verify the feasibility of kmer degeneration method. With a specific way of degenerating kmer, we could count and calculate the frequency of $K(N)$-mers. To eliminate the interference of random background, The frequency of random background should be subtracted by the frequency of K(N)-mers [10].

2.6.1. Frequency or Probability of Appearance of $(K)N$-mer. For a genome sequence of length L, we denote the frequency of appearance of the $(K)N$-mer $\alpha_1\alpha_2\cdots\alpha_N$ by $f(\alpha_1\alpha_2\cdots\alpha_N)$, where each $\alpha_i \in \{A, C, G, T\}$. This frequency divided by the total number $(L - K + 1)$ may be taken as the probability $p(\alpha_1\alpha_2\cdots\alpha_N)$:

$$p(\alpha_1\alpha_2\cdots\alpha_N) = \frac{f(\alpha_1\alpha_2\cdots\alpha_N)}{L - K + 1} \tag{8}$$

2.6.2. Subtraction of Random Background. According to Markov model prediction, we subtract the random background from the sequence as follows.

Suppose we have done $(K)N-1$-mer and $(K)N-2$-mer The probability of $(K)N$-mer is predicted by:

$$p^0(\alpha_1\alpha_2\cdots\alpha_N) = \frac{p(\alpha_1\alpha_2\cdots\alpha_{N-1})p(\alpha_2\alpha_3\cdots\alpha_N)}{p(\alpha_2\alpha_3\cdots\alpha_{N-1})} \tag{9}$$

In order to make sure all the frequencies of different kmer lengths (N, $N-1$ and $N-2$) in Equation 9 are non-zero, we should scan the genomes checking the different N values and determine the maximal N which makes all kinds of kmer appear at least once in every genome. For the data set in this study, we can get the maximal $N = 11$ (Table S4).

2.6.3. $(K)N$-mer vector and distance metric. We define:

$$a(\alpha_1\alpha_2\cdots\alpha_N) = \frac{p(\alpha_1\alpha_2\cdots\alpha_N) - p^0(\alpha_1\alpha_2\cdots\alpha_N)}{p^0(\alpha_1\alpha_2\cdots\alpha_N)}. \tag{10}$$

Then we can obtain the $(K)N$-mer vector for species A:

$$A = (a_1, a_2, \ldots, a_{4^N})$$

Likewise, we can get the $(K)N$-mer vector:

$$B = (b_1, b_2, \ldots, b_{4^N})$$

So the distance $D(A, B)$ between the two species is defined as

$$D(A,B) = \sqrt{\sum_{i=1}^{4^N}(a_i - b_i)^2} \tag{11}$$

2.6.4. Tree construction and decoration. MEGA, a popular software in molecular evolutionary genetic analysis, which could be employed to transform the distance matrix of species into phylogenetic tree by Neighbor Joining method. FigTree, which is designed as a graphical viewer of phylogenetic trees and as a program for producing publication-ready figures, could help us polish up the phylogenetic tree.

3. Results

3.1. The best N of $(K)N$-mer for a given K

As mentioned in the methods above, it is necessary to determine the best N value for a fixed K, since an N value that is too small will give a very noisy

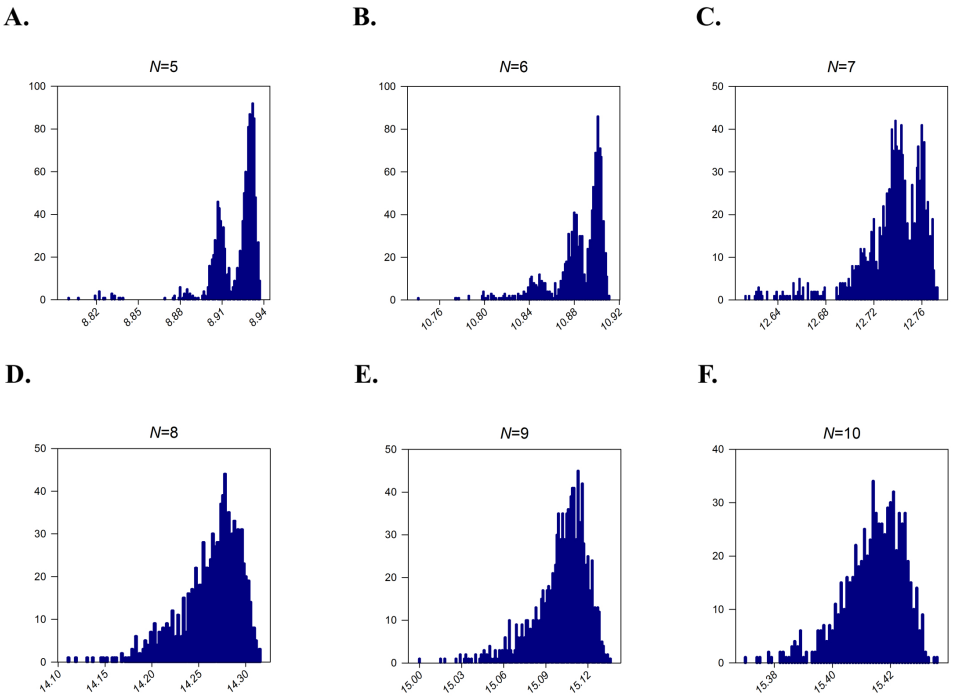


Figure 4: Determination of the optimal $K(N)$-mer for a distribution. Panel **A** to **F**. Distribution of the Shannon information of the 1000 randomly sampled combinations for $N = 5, 6, 7, 8, 9$ and 10, respectively. For small N values, the distributions were scattered with multiple peaks. Since $N = 8$, the distribution became one peak.

distribution, while too large will be very close to the K value and weaken the effectiveness of KD method. To assess the performance of different N values, we introduced a posterior distribution by picking 1000 combinations from $\binom{K}{N}$ and constructed a histogram of the Shannon information of 1000 combinations. We randomly selected a 50 Kbp genome sequence ($L = 50000$) and used $K = 50$ as an example to examine different N values. The use of this L length followed the previus work [7]. The results showed that for small N values, the distribution had multiple peaks. Since $N = 8$, the distribution became a single-peak distribution (Figure 4). We re-examined this for 10 different 50 Kbp sequences from human genomes and proved the robustness of the results. Therefore, the result of $N = 8$ was determined as a reference length of $(K)N$-mer.

3.2. Dispersed $(K)N$-mers have more information than continuous $(K)N$-mers

We focused on dispersed and continuous forms of combinations in $\binom{K}{N}$, which are defined in Section 2.3. We randomly selected 10 segment sequences from the human genome, as above, to demonstrate the result with the parameters of $L = 50000$(bp), $K = 50$(bp) and $N = 8$(bp). The results are shown in Table 1. It can be seen that all continuous $(K)N$-mers have Shannon information that is distributed extremely to the left side of the distribution. The majority is smaller than 5% quantile. By contrast, the dispersed $(K)N$-mers are distributed to the right side of the distribution and mostly larger than 95% quantile. Further, to compare the two forms' Shannon information on genome sequence and simulated random sequence, we generated a nucleotide sequence of 50kb and selected a 50kb segment sequence of human genome. Then we enumerated all forms of (20)5-mer and calculated their Shannon information. We found Shannon information of dispersed and continuous forms on simulated sequence was almost the same. However, for genome sequence, their information was significantly different. According to Table 2, we also found Shannon information of dispersed forms on genome sequence was significantly bigger than that of continuous forms.

3.3. Influence of K value for $(K)N$-mer in humans

It is necessary to assess the influence of various K values for a given N between dispersed and continuous $(K)N$-mers. We performed the assessment with 10 different genome sequence segments as repeats with $L = 50000$ and $N = 8$. For each sequence we considered 6 types of K values above $K = 50$. Table 3 indicates the result of one segment of sequence in the human genome. The results of all the 10 repeats of segment sequences are presented in Table S1. The results were similar to those in Table 1, which indicated that the difference between the dispersed and continuous form was not markedly influenced by K values less than 200. For K values greater than 200, the dispersed form of the $(K)N$-mer appeared more in the 50-75% interval. When the K values approached the level greater than 300, the dispersed form of the $(K)N - mer$ began to appear in the other side of the peak, in the interval of 25-50%.

One might be curious whether there is a definite upper limit of information volume for kmers. In fact, when the total kmer types were far greater than sequence length, all of the frequencies could be much closer to $\frac{1}{L}$; therefore, the upper limit of information for a given sequence length would be

Table 1: Difference of two (K)N-mer forms in the posterior distribution in human.

(K)N-mer	location in human genome		Quantiles of the posterior distribution					
form	chromosome	location	<5%	5%-25%	25%-50%	50%-75%	75%-95%	>95%
dispersed	chr13	79478999	0	0	0	1	15	34
dispersed	chr15	19839324	0	0	4	0	8	38
dispersed	chr20	41838329	0	0	0	0	32	18
dispersed	chr2	106032098	0	0	0	2	10	38
dispersed	chr2	220970069	0	0	0	1	11	38
dispersed	chr3	123809731	0	0	0	2	7	36
dispersed	chr6	148968552	0	0	0	4	16	30
dispersed	chr6	73531324	4	0	0	0	9	37
dispersed	chr8	125501152	0	0	0	2	23	25
dispersed	chr9	67772359	0	0	0	1	9	40
continuous	chr13	79478999	43	0	0	0	0	0
continuous	chr15	19839324	43	0	0	0	0	0
continuous	chr20	41838329	43	0	0	0	0	0
continuous	chr2	106032098	43	0	0	0	0	0
continuous	chr2	220970069	43	0	0	0	0	0
continuous	chr3	123809731	43	0	0	0	0	0
continuous	chr6	148968552	43	0	0	0	0	0
continuous	chr6	73531324	43	0	0	0	0	0
continuous	chr8	125501152	43	0	0	0	0	0
continuous	chr9	67772359	43	0	0	0	0	0

Notes: $L = 50000$, $K = 50$, $N = 8$.

Table 2: Difference of two forms on genome sequence and simulated sequence.

simulated sequence	<50%	>50%	<25%	>75%
continuous	9	7	0	1
dispersed	1	2	0	0
genome sequence	<50%	>50%	<25%	>75%
continuous	16	0	16	0
dispersed	0	3	0	3

Notes: $N = 5$, $K = 20$, $L = 50000$

$\log_2 L$. Notably, this value is independent from the K value. For the case of $L = 50000$ in this work, the upper limit of information I_{max} is

$$I_{max} = \log_2 L = \log_2 50000 = 15.6096 \tag{12}$$

The $(K)N$-mers were compared with $N = 8$ and various K values (Figure 5). It could be seen that the information increased slightly for kmers with

Table 3: Difference of two (K)N-mer forms under different K values.

	kmer length	Quantiles of the posterior distribution					
(K)N-mer form	(K value)	<5%	5%-25%	25%-50%	50%-75%	75%-95%	>95%
dispersed	30	0	0	0	0	5	45
dispersed	50	0	0	0	1	15	34
dispersed	70	0	0	0	2	22	26
dispersed	100	0	0	0	1	17	32
dispersed	150	0	0	0	0	15	35
dispersed	200	0	0	0	1	13	36
dispersed	220	0	0	0	7	24	19
dispersed	240	0	0	0	4	12	34
dispersed	260	0	0	0	2	9	39
dispersed	280	0	0	1	3	25	21
dispersed	300	0	0	1	3	26	20
dispersed	400	0	0	0	2	19	29
dispersed	500	0	0	1	12	21	16
continuous	30	23	0	0	0	0	0
continuous	50	43	0	0	0	0	0
continuous	70	43	0	0	0	0	0
continuous	100	43	0	0	0	0	0
continuous	150	43	0	0	0	0	0
continuous	200	43	0	0	0	0	0
continuous	220	43	0	0	0	0	0
continuous	240	43	0	0	0	0	0
continuous	260	43	0	0	0	0	0
continuous	280	43	0	0	0	0	0
continuous	300	43	0	0	0	0	0
continuous	400	43	0	0	0	0	0
continuous	500	43	0	0	0	0	0

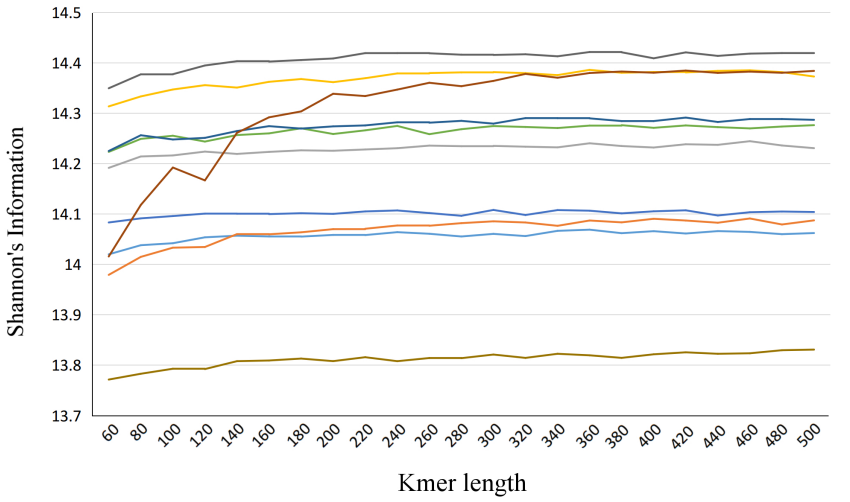
Notes: $K = 50$, $N = 8$, human chromosome 13, start position at 79478999.

lengths less than 200. For kmers longer than 200, the information tended to fluctuate. It was noticeable that the difference of Shannon information is significant in different genome segment sequences, which can reflect heterogeneity of the genome sequences.

3.4. Comparison of different $(K)N$-mer combinations in pan-mammals

We examined the difference between dispersed and continuous $(K)N$-mers with different mammalian genomes. Species were selected from the main branches of mammals. It can be seen from Table 4 (one segment of sequence from one species) and Table S2 (for all other sequences) that for all genome

A.



B.

K-mer Length	Positions for *N*-mer
60	1 – 9 – 17 – 25 – 33 – 41 – 49 – 60
80	1 – 12 – 23 – 34 – 45 – 56 – 67 – 80
100	1 – 15 – 29 – 43 – 57 – 71 – 85 – 100
120	1 – 18 – 35 – 52 – 69 – 86 – 103 – 120
140	1 – 20 – 39 – 58 – 77 – 96 – 115 – 140
160	1 – 23 – 45 – 67 – 89 – 111 – 133 – 160
180	1 – 26 – 51 – 76 – 101 – 126 – 151 – 180
200	1 – 29 – 57 – 85 – 113 – 141 – 169 – 200
220	1 – 32 – 63 – 94 – 125 – 156 – 187 – 220
240	1 – 35 – 69 – 103 – 137 – 171 – 205 – 240
260	1 – 38 – 75 – 112 – 149 – 186 – 223 – 260
280	1 – 40 – 79 – 118 – 157 – 196 – 235 – 280
300	1 – 43 – 85 – 127 – 169 – 211 – 253 – 300
320	1 – 46 – 91 – 136 – 181 – 226 – 271 – 320
340	1 – 49 – 97 – 145 – 193 – 241 – 289 – 340
360	1 – 52 – 103 – 154 – 205 – 256 – 307 – 360
380	1 – 55 – 109 – 163 – 217 – 271 – 325 – 380
400	1 – 58 – 115 – 172 – 229 – 286 – 343 – 400
420	1 – 60 – 119 – 178 – 237 – 296 – 355 – 420
440	1 – 63 – 125 – 187 – 249 – 311 – 373 – 440
460	1 – 64 – 131 – 196 – 261 – 326 – 391 – 460
480	1 – 69 – 137 – 205 – 273 – 341 – 409 – 480
500	1 – 72 – 143 – 214 – 285 – 356 – 427 – 500

Figure 5: Influence of the K value on the absolute value of Shannon information for the dispersed $K(N)$-mer form. **A**. Changes in the Shannon information value with K values for 10 different human genome segment sequences. Each color denotes one sequence. **B**. The positions chosen from the K-mer for the $K(N)$-mer. The K-mer lengths corresponded to those in panel **A**.

sequences from the 5 mammalian species, the information values from continuous combinations gathered far away to the left side of the distribution

peak, while the values from dispersed combinations were located in the right side of the distribution peak.

Table 4: Difference of two (K)N-mer forms in difference mammals.

Species	Sequence location	K(N)-mer form	Quantiles of the posterior distribution					
			<5%	5%-25%	25%-50%	50%-75%	75%-95%	>95%
mouse	Chr11.118185247	dispersed	0	0	0	0	15	35
dog	Chr10.19378736	dispersed	0	0	0	0	5	45
elephant	Scaffold12.49199351	dispersed	0	0	0	1	22	27
Wallaby	Scaffold1097.3543	dispersed	0	0	0	3	34	13
mouse	Chr11.118185247	continuous	43	0	0	0	0	0
dog	Chr10.19378736	continuous	43	0	0	0	0	0
elephant	Scaffold12.49199351	continuous	43	0	0	0	0	0
Wallaby	Scaffold1097.3543	continuous	43	0	0	0	0	0

Notes: $K = 50$, $N = 8$, the chromosome/scaffold and the start position were divided by a dot.

3.5. Assessment of kmer degeneration method in tree construction

To test the availability of KD method in alignment-free phylogenomics approach, we constructed the phylogenic relation of *Carnivora* with 16 genomes as a case. The data set included 14 carnivorous species as well as human and mouse genome as outgroup. The N value was determined as 11 and we checked a series of K values of 50, 60, 70, 80, 90, 100, 110, 120, 130, 200, 250 and 300. The results showed that when K =50, 60, 70, 80, 90, 110, the topology of Canine suborder tree is the same with the general accepted phylogeny (Fig. 6A) [8]. What should be noticed is the position of the *Pinnipedia*. In some studies, it is sister group of the Weasel superfamily, while in another studies, it is sister group of it is sister group of *Arctoidea* superfamily [1, 9]. Our result suggests that it is sister group of the Weasel superfamily instead of the *Arctoidea* superfamily [8]. There are two kinds of topology of tree for the *Feliformia* suborder in our result. When K =50, 70, 80, 90 and 110, cheetah is clustered with the linage of *Panthera* genous while domestic cat and puma are clustered as one lineage. While in the case of $K = 100$, none of the three species, cheetah, puma and domestic cat, have been clustered (Fig. 6B). Compared with our alignment-free result, it has been generally accepted that the three species should be clustered into one monophyletic group (Fig. 6D) [6, 8]. Despite of this discrepancy, our result is better than those using non-degenerated kmer with the same N value (Fig. 6C).

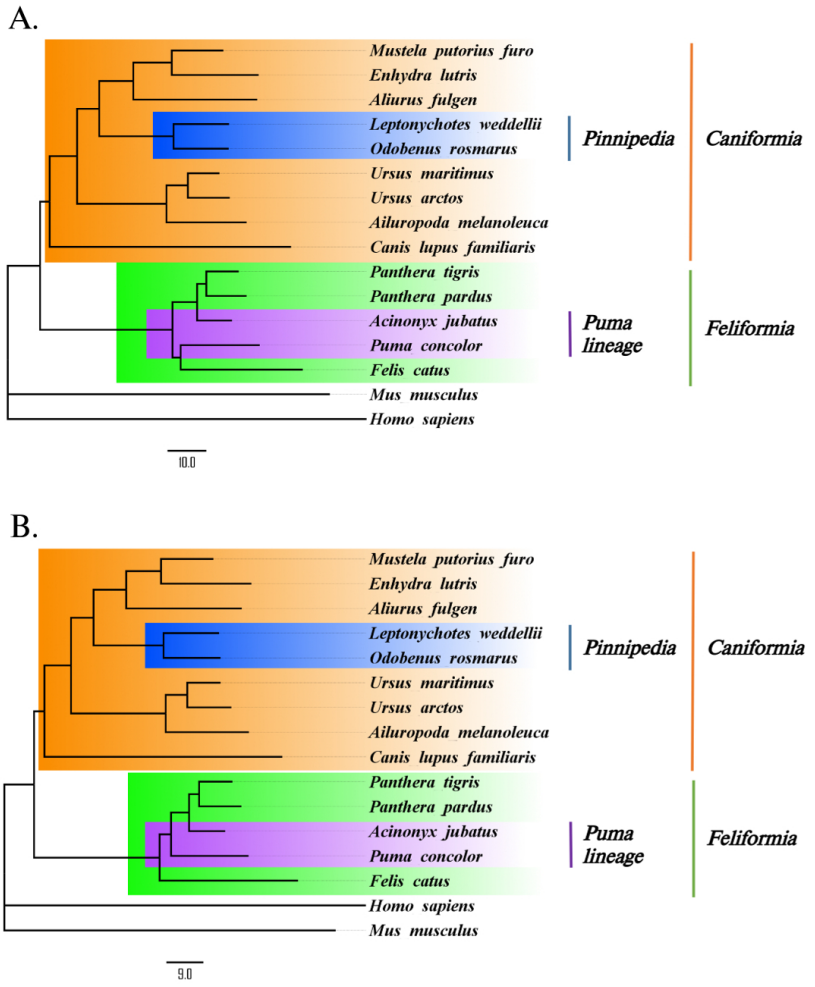
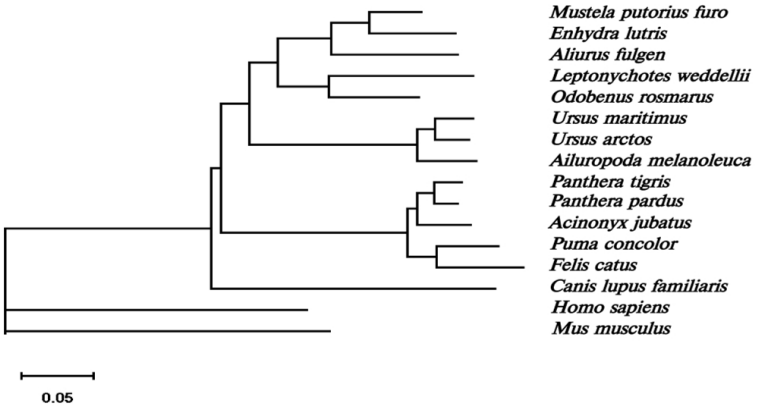


Figure 6: Performance of kmer degeneration employed for improving cvTree constructing phylogenetic tree of complicated mammal genomes. **A**. phylogenetic tree of $K = 50$; **B**. phylogenetic tree of $K = 60$;

4. Conclusions

The aim of this study was to enable the alignment-free whole-genome phylogeny for complicated large genomes, such as in mammals or birds. For the application of alignment-free methods, the large genome size confined the usage of long K-mers which have a vast amount of kmer types. Our KD method reduced the K-mer types while maintaining a long span as K-mer.

C.



D.

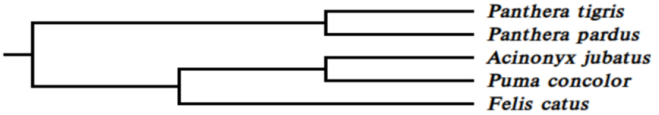


Figure 6: **C**. phylogenetic tree by cvTree; **D**. *Feliformia* suborder referred from [6, 8].

This makes it possible to investigate whole-genome phylogeny with frequencies of long K-mers. For a given K, we first use the distribution of Shannon information to determine the best N for $(K)N$-mer. Then by analysis of different combinations of positions, we find that the cases of continuous N positions possess significantly less information than the cases of dispersed N positions for a fixed K. So for a DNA sequence, one can use dispersed $(K)N$-mer to substitute K-mer. We applied our method to Carnivora with 16 genomes, which is better than non-degenerated kmer with the same N value. Therefore, our method can be used to other genomes for future research.

5. Conflict of interest statement

The authors declare no competing financial interests.

Acknowledgements

This work is supported by National Natural Science Foundation of China grant (#91746119) and the Strategic Priority Program of the Chineses Academy of Science (XBD31020000).

References

[1] U. Arnason, A. Gullberg, and M. Janke, Akullberg, *Mitogenomic analyses of caniform relationships*, Molecular Phylogenetics & Evolution **45** (2007), no. 3, 863–874.

[2] G. Bernard, C. X. Chan, Y. B. Chan, X. Y. Chua, Y. Cong, J. M. Hogan, S. R. Maetschke, and M. A. Ragan, *Alignment-free inference of hierarchical and reticulate phylogenomic relationships*, Briefings in Bioinformatics (2017).

[3] B. E. Blaisdell, *A measure of the similarity of sets of sequences not requiring sequence alignment*, Proceedings of the National Academy of Sciences of the United States of America **83** (1986), no. 14, 5155–5159.

[4] C. X. Chan, G. Bernard, O. Poirion, J. M. Hogan, and M. A. Ragan, *Inferring phylogenies of evolving sequences without multiple sequence alignment*, Sci. Rep. **4** (2014), no. 39, 6504.

[5] M. Deng, C. Yu, Q. Liang, R. L. He, and S. S. T. Yau, *Correction: A novel method of characterizing genetic sequences: Genome space with biological distance and applications*, Plos One **6** (2011), no. 3, e17293.

[6] G. Li, B. Davis, E. Eizirik, and W. Murphy, *Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae)*, Genome Research **26** (2015), no. 1, 1.

[7] A. J. Gentles and S. Karlin, *Genome-scale compositional comparisons in eukaryotes*, Genome Research **11** (2001), no. 4, 540–546.

[8] W. E. Johnson, E. Eizirik, J. Pecon-Slattery, W. J. Murphy, A. Antunes, E. Teeling, and S. J. O'Brien, *The late Miocene radiation of modern Felidae: a genetic assessment.*, Science **311** (2006), no. 5757, 73–77.

[9] R. Peng, B. Zeng, X. Meng, B. Yue, Z. Zhang, and F. Zou, *The complete mitochondrial genome and phylogenetic analysis of the giant panda (Ailuropoda melanoleuca)*, Gene **397** (2007), no. 1, 76–83.

[10] J. Qi, B. Wang, and B. I. Hao, *Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach*, Journal of Molecular Evolution **58** (2004), no. 1, 1.

[11] C. E. Shannon, *A mathematical theory of communication*, Bell Labs Technical Journal **27** (1948), no. 3, 379–423.

[12] 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), no. 8, 2677–2682.

[13] Z. Xu and B. Hao, *CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes*, Nucleic Acids Research **37** (2009), no. Web Server issue, W174–W178.

[14] A. Zielezinski, S. Vinga, J. Almeida, and W. M. Karlowski, *Alignment-free sequence comparison: benefits, applications, and tools*, Genome Biology **18** (2017), no. 1, 186.

Shuai Liu[2,3], Shaojun Pei[1], Stephen S.-T. Yau[1,*], and Qi Wu[3,4,*]

1. Department of Mathematical Sciences, Tsinghua University
Beijing 100084, China

2. Institute of Physical Science and Information Technology
Anhui University, Hefei 230601, China

3. Key Laboratory of Animal Ecology and Conservation Biology
Institute of Zoology, Chinese Academy of Sciences, Beijing, China

4. The State Key Laboratory of Mycology (SKLM)
Institute of Microbiology, Chinese Academy of Science, Beijing, China

∗ Corresponding authors.
E-mail address: yau@uic.edu (Stephen S.-T. Yau)
E-mail address: ribozyme@ioz.ac.cn (Qi Wu)