Contents lists available at SciVerse ScienceDirect

Gene

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

Segmented *K*-mer and its application on similarity analysis of mitochondrial genome sequences

Hong-Jie Yu *

Department of Mathematics, School of Science, Anhui Science and Technology University, Fengyang, Anhui 233100, China
Intelligent Computing Laboratory, Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, P.O. Box 1130, Hefei, Anhui 230031, China
Department of Automation, University of Science and Technology of China, Hefei, China

ARTICLE INFO

Article history:
Accepted 19 December 2012
Available online xxxx

Keywords:
Segmented *K*-mer
Sparseness degree
Similarity analysis
Mitochondrial genome
Phylogenetic tree

ABSTRACT

K-mer-based approach has been widely used in similarity analyses so as to discover similarity/dissimilarity among different biological sequences. In this study, we have improved the traditional *K*-mer method, and introduce a segmented *K*-mer approach (*s*-*K*-mer). After each primary sequence is divided into several segments, we simultaneously transform all these segments into corresponding *K*-mer-based vectors. In this approach, it is vital how to determine the optimal combination of distance metric with the number of K and the number of segments, i.e., (K^*, s^*, and d^*). Based on the cascaded feature vectors transformed from s^* segmented sequences, we analyze 34 mammalian genome sequences using the proposed *s*-*K*-mer approach. Meanwhile, we compare the results of *s*-*K*-mer with those of traditional *K*-mer. The contrastive analysis results demonstrate that *s*-*K*-mer approach outperforms the traditionally *K*-mer method on similarity analysis among different species.

1. Introduction

Sequence comparison is the most basic task in the field of computational molecular biology that arises in evolutionary, structural or functional studies of biological sequences, such as DNA and protein sequences. The aim of sequence comparison is to discover similarity relationships between various biological sequences. Several efficient methods have been proposed to process and analyze these sequences, where each DNA sequence is regarded as a string over the 4-letter alphabet {Ω = A, G, C, T}. Generally, these methods can be categorized into two classes: alignment-based and alignment-free. However, for both the two methods, their common problems are how to efficiently extract essential information from the sequences.

There are many approaches for efficiently transforming DNA sequences into numerical signals. Generally, one can use binary sequences to describe the position of each symbol (Voss, 1992). Also, several other different transformation methods have been proposed (Akhtar et al., 2007; Brodzik and Peters, 2005; Cristea, 2003; Jeffrey, 1990; Liao et al., 2005; Randić, 2008; Wang et al., 2009; Zhang and Zhang, 1994). Moreover, many frequency-based algorithms have been introduced for sequence comparisons as indicated in Jun et al. (2009), Sims, Jun, et al. (2009a,b) and Wu et al. (1997, 2001, 2005). *K*-mer, as one of approaches to sequence comparison, is widely used in similarity analysis, where a DNA sequence is summarized by *l*-tuples consisting of all *K*-mer counts, $K = 2, 3, \ldots, 9$ usually (Karlin and Burge, 1995). Thus, DNA sequences with different lengths can be uniformly transformed into equal-length vectors.

Compared to the standard approach to building evolutionary tree using multiple sequence alignments (MSA), the advantage of *K*-mer analysis is that the word frequencies-based approach is much faster, and may therefore be used for comparison of whole genomes. However, a deficiency is the loss of information since the huge amount of DNA sequence data is condensed into a vector of *K*-mer counts. Moreover, another problem is that the order of *K*-mers in compared sequences is more or less neglected.

In addition, for two genome sequences, there are both whole similarity and local similarity. So, it is not appropriate that only the whole similarity is focused on when we analyze similarity of sequence. In this study, under the framework of optimization, we propose a novel improved *K*-mer-based approach for similarity analyses among different genome sequences, where each primary genome sequence is divided into several segments and each segment is transformed into a vector via traditional *K*-mer. Thus a cascaded vector is obtained though the proposed approach. The procedure consists of three steps: a) Exploring the combination of optimal distance

Abbreviations: A, adenosine; bp, base pair(s); C, cytidine; CV, composition vector; FV, feature vector; G, guanosine; mt, mitochondria; MEGA, molecular evolutionary genetics analysis; MSA, multiple sequence alignments; nt, nucleotide(s); Pdist, pair-wise distance; *s*-*K*-mer, segmented *K*-mer; T, thymidine.

* Department of Mathematics, School of Science, Anhui Science and Technology University, Fengyang, Anhui 233100, China. Tel.: +86 550 6732393; Fax: +86 0550 6733 165.

E-mail address: yhj551@gmail.com.

0378-1119/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2012.12.079

metric with the optimal number K^* of K-mer for genome sequence dataset; b) searching the optimal number of segments s^* for several cascaded 4^{K^*} dimensional vector; and c) analyzing their similarities via a phylogenetic tree based on the obtained cascaded vector. The validity of the proposed approach is demonstrated via its application on real dataset.

2. Descriptors for genome sequences

To describe the biological sequences, many authors designed descriptors that can be used as the components of similarity measures among multiple sequences (Bielińska-Wąż, 2011; Liao and Wang, 2004a,b; Randić, 2004, 2008; Randić et al., 2003a,b,c; Song and Tang, 2005; Yang et al., 2012; Yao and Wang, 2004). In this section, we propose novel descriptors to characterize each genome sequence, and apply descriptors on the similarity analysis of sequences.

2.1. K-mer of sequences

Hao Bailin's laboratory developed a K-mer-based composition vector (CV) method with subtracted background 'noise' modeled by a Markov chain estimator. Using this K-mer-based method, Hao's group obtained valuable results for both protein and genome sequences (Gao and Qi, 2007; Qi et al., 2004). Likewise, a general description of the FFP method has been published (Sims et al., 2009a,b).

For the K-mer method, a description of the details is given as follows.

Let $\vec{\mathbf{N}}$ be the primary genome sequence with length L, $\vec{\mathbf{N}} = {}'\mathbf{N}_1\mathbf{N}_2 \cdots \mathbf{N}_L{}'$, where $\mathbf{N}_i \in \{A,T,G,C\}$.

A K-mer is a series of K consecutive letters in a sequence. The standard approach for counting K-mers in a sequence of length L uses a sliding window of length K, shifting the frame one base each time from position 1 to $L-K+1$, until the entire genome is scanned. The derived feature vector can be indicated as

$$\vec{\mathbf{F}} = (f_1, f_2, \cdots, f_{4^K}) \tag{1}$$

where f_i is the raw frequency of the corresponding feature and $N = 4^K$ is the total number of all possible K-mers. Then, $\vec{\mathbf{F}}$ can be normalized by the length $L-K+1$, which frees from the influence of different lengths of each genome sequence.

2.2. Segmented K-mer of sequences

2.2.1. The optimal number K^* for K-mer

For a given group of genome sequences, we need to devise a criterion for the determination of the optimal number K^*.

In general, to validate a newly proposed improved algorithm, one can compare the results of the improved approach with that of the traditional one. In order to make a comparison, a correlation analysis is provided. We calculate the pair-wise distances of the group of genome sequences using MEGA software based on alignment framework. Here, the obtained alignment-based data is just used as a reference system, which can be used for quantitatively comparing the performance for the improved K-mer with that one of the traditional K-mer.

The optimal number K^* can be determined by:

$$(K^*, \theta^*) = \underset{K,\theta}{argmax}\ corr_{K,\theta}(Pdist_K, Pdist_0) \tag{2}$$

where K^* indicates the optimal number of K-mer, while the θ^* denotes the optimal distance metric at this time, and $\theta^* \in \Theta$, $\Theta = \{$'euclidean', 'cityblock', 'minkowski', 'cosine', 'correlation', 'spearman'$\}$. At the right side of Formula (2), there are two input parameters, where $Pdist_K$ stands for the pair-wise distance vector based on K-mer, while $Pdist_0$ denotes the pair-wise distance vector via the alignment-based method.

2.2.2. Considering local similarity

Under the frame of alignment-based method, the Needleman–Wunsch algorithm investigates the comparison of the whole length of two sequences and therefore performs a global optimal alignment. However, it is also important to find a local similarity among sequences in many cases (Xu and Wunsch, 2005). Thus, it is vital that the local similarity for the multiple sequences should be considered in order to improve the precision of similarity analysis of sequence, when the K-mer approach is used.

2.3. The optimal segmentation scheme

After the optimal number K^* for K-mer has been determined, one can furthermore search the optimal number of segment for the segmented K-mer.

Let s be the number of segment for a certain segmentation scheme, where $s = 2, 3, \ldots, M$, and M is the maximal number of segments. For a given sequence with the length L, one can calculate the mod for L/s, which is denoted as:

$$m = mod(L, s) \tag{3}$$

Then the first segment of the primary sequence $\vec{\mathbf{N}}^{(1)}$ can be correspondingly transformed into the first 4^K dimensional segmented feature vector $\vec{\mathbf{F}}^{(1)}$, where $\vec{\mathbf{N}}^{(1)}$ is extracted from the primary sequence, ranging from the 1st to the mth locus, while the $\vec{\mathbf{F}}^{(1)}$ denotes the first transformed vector via the traditional K-mer method.

By analogy, the second segment of the primary sequence $\vec{\mathbf{N}}^{(2)}$, ranging from the $(m+1)$th to $(2\ m)$th locus, can be correspondingly transformed into the second 4^K dimensional segmented feature vector $\vec{\mathbf{F}}^{(2)}$, and so on. Finally, the last segment $\vec{\mathbf{N}}^{(s)}$ from the primary sequence, ranging from the $((s-1)*m+1)$th locus to the end, can be transformed into the last 4^K dimensional segmented feature vector $\vec{\mathbf{F}}^{(s)}$.

Thus, for a preset integer s, each genome sequence $\vec{\mathbf{N}}$ can be uniformly transformed into a corresponding $s \times 4^{K^*}$ dimensional feature vector, denoted as $\tilde{\mathbf{F}}$, where

$$\tilde{\mathbf{F}} = \left(\vec{\mathbf{F}}^{(1)}, \vec{\mathbf{F}}^{(2)}, \cdots, \vec{\mathbf{F}}^{(s)}\right) \tag{4}$$

while $s = 2, 3, \ldots, M$, and M is the maximal number of segments. All $\vec{\mathbf{F}}^{(s)}$ are calculated via Formula (1), and K^* is determined by Formula (2).

2.4. Compound s-K-mer

Thus, based on the obtained optimal K^*, the optimal number for segmentation s^* can be determined by:

$$\left(s^*, \tilde{\theta}^*\right) = \underset{s,\tilde{\theta}}{argmax}\ corr_{s,\tilde{\theta}}(Pdist_{sK^*}, Pdist_0) \tag{5}$$

where s^* denotes the optimal number of segmentation for the compound s-K-mer, and the $\tilde{\theta}^*$ indicates the updated optimal distance metric for the right side of Formula (5), where the $Pdist_{sK^*}$ stands for the pair-wise distance among genome sequences via the improved feature vector $\tilde{\mathbf{F}}$ based on Formula (4).

Please cite this article as: Yu, H.-J., Segmented K-mer and its application on similarity analysis of mitochondrial genome sequences, Gene (2013), http://dx.doi.org/10.1016/j.gene.2012.12.079

The proposed segmented *K*-mer algorithm can be summarized as follows.

```
Input: multiple nucleotide sequences: $N^{(1)}, N^{(2)}, \ldots, N^{(n)}$
begin
    for $k = 2$ to $K$ do
        for $l = 1$ to $n$ do
            Transform each sequence $N^{(l)}$ into $4^k$ dimensional
                feature vector $F^{(l)}$ by traditional K-mer method
                via Formula (1)
        end for
    end for
    Explore the optimal $K^*$ and $\theta^*$ through Formula (2)
    for $s = 2$ to $M$ do
        Divide the sequences into $s$ segments according to the
            scheme described in Section 2.3
        Transform each sequence $N^{(l)}$ into $s*4^{K^*}$ dimensional fea-
            ture vector $F^{(l)}$ by improved K-mer approach, i.e., seg-
            mented K-mer, see Formula (4) for details
        Search the optimal $s^*$ and $\tilde{\theta}*$ through Formula (5)
    end for
    Draw the dendrogram using the pair-wise distances matrix
        based on $K^*$ and the output of Formulas (4) and (5)
end
```

3. Application upon real dataset

In this section, we apply the proposed *s*-*K*-mer approach upon the real genome dataset, i.e. the mitochondrial genome dataset. This dataset can be directly obtained from Huang et al. (2011) and Yu et al. (2010). The MATLAB source code of our proposed method can be downloaded at:http://home.ustc.edu.cn/~yhj70/sKmer/code.rar.

3.1. Preparation for optimization

The accession numbers of these 34 species in the GenBank are as follows: Human, V00662; Common Chimpanzee, D38113; Gorilla, D38114; Pigmy Chimpanzee, D38116; Gibbon, X99256; Baboon, Y18001; Vervet Monkey, AY863426; Ape, NC_002764; Sumatran Orangutan, NC_002083; Bornean Orangutan, D38115; Cat, U20753; Pig, AJ002189; Sheep, AF010406; Goat, AF533441; Cow, V00654; Buffalo, AY488491; Dog, U96639; Wolf, EU442884; Leopard, EF551002; Tiger, EF551003; White Rhinoceros, Y07726; Indian Rhinoceros, X97336; Harbor Seal, X63726; Gray Seal, X72004; African Elephant, AJ224821; Asiatic Elephant, DQ316068; Brown Bear, AF303110; Polar Bear, AF303111; Black Bear, DQ402478; Rabbit, AJ001588; Squirrel, AJ238588; Hedgehog, X88898; Vole, AF348082; Norway Rat, X14848.

We calculate the pair-wise distances of these 34 sequences using MEGA software based on the alignment framework (Saitou and Nei, 1987). The alignment-based results of the pair-wise distances are listed in Table S1 (See at: http://home.ustc.edu.cn/~yhj70/sKmer/Pdist_0_34.xls), from which we can extract 33 entries on the first row comprising of the distances between Human and the rest 33 species. In addition, the correlation degree between every two different results from each approach is an effective measure to determine whether a new approach is effective or not. The higher correlation degree with the traditional alignment-based method means that the new approach is effective.

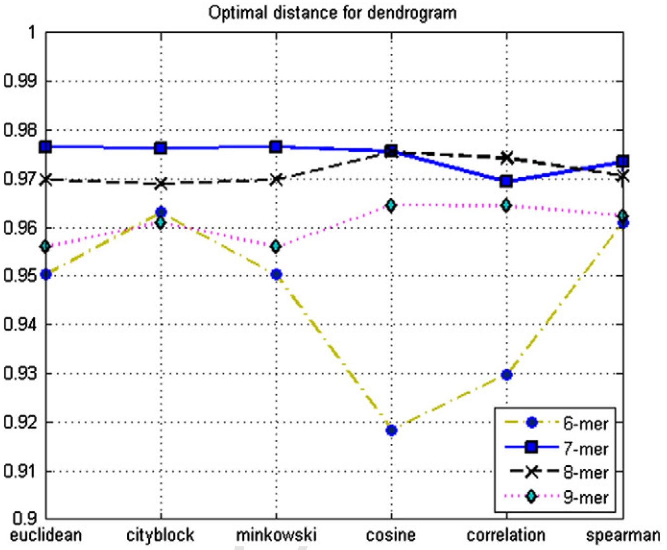

Fig. 1. Explore the optimal type of distance metric via the performance values against all the six kinds of distance metric. The *y*-axis indicates that the correlation coefficients between the pair-wise distance vector from *K*-mer and that one from the traditional alignment-based approach via MEGA software. Only the results of $K = 6$, 7, 8, and 9 are shown.

3.2. Optimizing K^* for K-mer

To explore which distance metric θ^* is optimal and how many K^* is the optimal, we selected six kinds of distance metric upon *K*-mer, $K = 2, 3, \ldots, 9$, respectively. According to the procedure for *s*-*K*-mer algorithm depicted in Section 2.4, we calculate all the eight performances vs. their corresponding *K*-mer, $K = 2$, 3, …, 9, respectively. To compare with the traditional alignment-based method, we calculate the correlation degrees for the results of our approach from different distance metrics. The performance values are shown in Fig. 1. Through Formula (2), we need determine the optimal distance metric via the performance values against all the six kinds of distance metric. As shown in Fig. 1, it can be seen that the performance values from 7-mer are robust and are mostly greater than those from other cases, such as 6-mer, 8-mer, 9-mer etc. Thus, through Formula (2), it can be determined that the optimal distance metric is 'euclidean'.

Moreover, under the circumstance in which $\theta^* =$ 'euclidean', we can also obtain that the optimal number for *K*-mer just equals to 7, i.e. $K^* = 7$. The more detailed results are shown in Fig. 2, where it

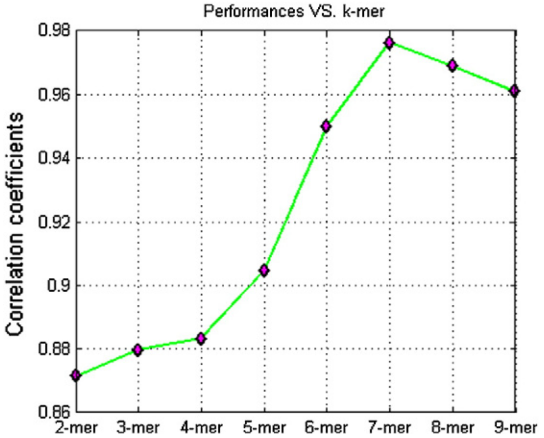

Fig. 2. Explore the optimal number K^* for *K*-mer under the circumstance of the explored optimal distance metric shown in Fig. 1, i.e. $\theta^* =$ 'euclidean'. Similarly, the *y*-axis denotes the performance values upon all the *K*-mer, where *K* ranges from 2 to 9, respectively. (See the caption of Fig. 1).

Please cite this article as: Yu, H.-J., Segmented *K*-mer and its application on similarity analysis of mitochondrial genome sequences, Gene (2013), http://dx.doi.org/10.1016/j.gene.2012.12.079

can be found that the performance firstly increases then decreases when K changes from 6 to 9. In particular, the value of performance reaches a peak at the point of 7-mer, where the peak value equals to 0.9763.

For a DNA sequence with average length 17,000 bps, the K-mer count vector $\vec{\mathbf{F}}$ becomes too sparse for $K \geq 8$. Thus, the comparisons among long sequences based on the higher order K-mer may not capture the essential feature of sequences. Therefore, we need consider ranking of K-mers in terms of their sparseness degree. To manifest the latent reason for the explored optimal number K^* of K-mer, we calculate the sparseness degrees for all the 34 species based on K-mer via the Formula (6) as follows:

$$sp = \frac{n_0}{4^K} \tag{6}$$

where sp denotes the sparseness degrees for the 4^K dimensional feature vector, while n_0 indicates the total number of zero elements. Obviously, for $n_0 < 4^K$ and $n_0 \geq 0$ always hold, sp belongs to an interval $[0, 1)$.

The results are shown in Fig. 3, which demonstrates that the sp values for 5-mer and 6-mer are all lower, while those sp values from 8-mer and 9-mer are all greater than 80%. However, only the sp values for 7-mer are moderate, around 50%. Moreover, Fig. 3 shows that the sp values both from 7-mer and from 6-mer vary violently, which may contribute to the distinguishable ability for K-mer among these 34 species. However, the K-mer's distinguishable ability may decrease when K increases to 8 or 9, for these two curves become more flat.

Figs. 1, 2, and 3 illustrate that the K-mer's distinguishable ability will increase firstly and then decrease. Fig. 2 shows that the performance has marked peak value against K-mer. Subsequently, in order to overcome the limitation of traditional K-mer method, we will further improve the performance on similarity analysis via the proposed segmented K-mer approach, i.e. s-K-mer.

3.3. The performance for s-K-mer approach

As described in Section 2.4, where we have investigated an improvement on traditional K-mer method, we apply it upon the above-mentioned data set.

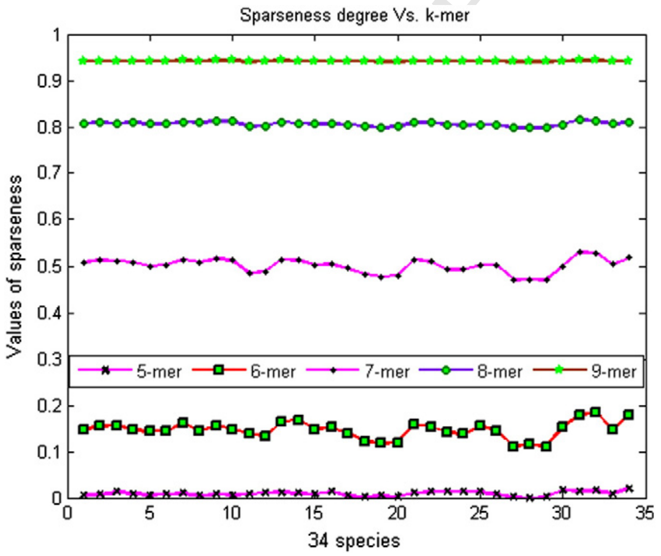

Fig. 3. Sparseness degrees for all 34 species based on their feature vector via K-mer. Only the results of $K = 5$, 6, 7, 8, and 9 are shown. The x-axis indicates the 34 vectors from each different species, while the y-axis denotes the sp values for those 5 kinds of K-mer, respectively.

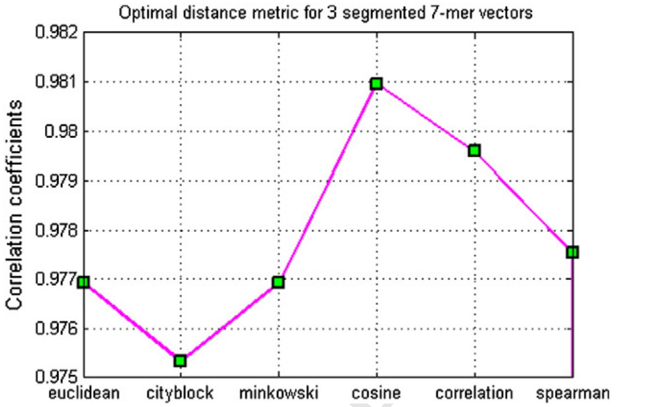

Fig. 4. Under the circumstance of $(K^*, \theta^*) = (7,$ 'euclidean'), the explored optimal number of segmented K-mer $s^* = 3$. The performance values against all the six kinds of distance metric, respectively. Now, the subsequently explored optimal distance metric $\tilde{\theta}^*$ is 'cosine'. Similarly, the y-axis denotes the performance values. (See the caption of Fig. 1 for details).

According to Formula (5), it can be determined that the optimal number of segmented K-mer is 3, i.e. $s^* = 3$. Meanwhile, $\tilde{\theta}^*$ is 'cosine' rather than 'euclidean'. The results are shown in Fig. 4, where it can be seen that the performance value reaches to 0.981 at peak point. Compare Fig. 4 with Fig. 2, we can find that s-K-mer method outperforms that one from the traditional K-mer.

To observe the change of sparseness degree for K-mer, $K = 5$, 6, 7, 8, and 9, we list the sparseness value for all these 34 species through five kinds of K-mer, respectively. Meanwhile, we also calculate the sparseness degree values via 3 segmented 7-mer. The results are listed in Table 1, where the fourth column is the entries for the optimal K-mer, i.e. $K^* = 7$, while those entries for 3 segmented 7-mer are listed in the last column, respectively.

In general, the statistic index of σ/μ can reflect the discrete degree of a group data. Specifically, the values of σ/μ from sparseness degree listed in Table 1 can more or less indicate the distinguishable ability that which K-mer is better. Therefore, σ/μ can be served as performance index to determine the optimal K-mer. However, the value of σ/μ is just a necessary but not sufficient condition for distinguishing species. Then, we calculate the mean values and standard deviation values for all the six kinds of K-mers listed in Table 1.

The results for performance index σ/μ are listed at the bottom of Table 1, from which we can see that the value of σ/μ from the 3 segmented 7-mer keeps close to that one from 7-mer. In fact, Table 1 shows that the former one and the later one are 0.0415 and 0.0316, respectively. Compare with those ones in 4th column, all the 34 entries in the last column indicate that the sparseness degrees from 3 segmented 7-mer not only become smaller than those ones from 7-mer, but also the value of σ/μ keeps still close to each other. This phenomenon indicates that 3 segmented 7-mer is effective to a certain extent.

3.4. Phylogenetic analysis of genomes via s-K-mer

In general, the phylogenetic analysis among genome sequences can be performed in the following steps:

(1) Firstly, we calculate each feature vector (FV) for each genome sequence based on optimal s-K-mer;
(2) Secondly, we explore the upgraded optimal distance metric, through which we can calculate pair-wise distance matrix;
(3) Finally, we investigate the optimal 'linkage' for plotting dendrogram.

Please cite this article as: Yu, H.-J., Segmented K-mer and its application on similarity analysis of mitochondrial genome sequences, Gene (2013), http://dx.doi.org/10.1016/j.gene.2012.12.079

Table 1
Sparseness degree for the *K*-mer and 3-segmented-7-mer.

Species\K-mer	5-mer	6-mer	7-mer	8-mer	9-mer	3–7-mer
Human	0.0059	0.1460	**0.5075**	0.8078	0.9422	*0.1027*
Pigmy_chimpanzee	0.0068	0.1555	**0.5123**	0.8090	0.9425	*0.1043*
Common_chimpanzee	0.0127	0.1560	**0.5103**	0.8078	0.9422	*0.1038*
Gorilla	0.0088	0.1467	**0.5074**	0.8087	0.9427	*0.1017*
Gibbon	0.0059	0.1436	**0.4988**	0.8059	0.9423	*0.1023*
Baboon	0.0078	0.1436	**0.5021**	0.8062	0.9421	*0.1020*
Cercopithecus_aethiops	0.0098	0.1621	**0.5141**	0.8102	0.9430	*0.1070*
Ape	0.0039	0.1433	**0.5077**	0.8091	0.9426	*0.1035*
Bornean_orangutan	0.0078	0.1563	**0.5151**	0.8128	0.9435	*0.1043*
Sumatran_orangutan	0.0059	0.1475	**0.5120**	0.8121	0.9433	*0.1044*
Cat	0.0078	0.1392	**0.4839**	0.8007	0.9413	*0.0992*
Dog	0.0107	0.1318	**0.4872**	0.8024	0.9417	*0.0990*
Pig	0.0117	0.1646	**0.5139**	0.8106	0.9429	*0.1067*
Sheep	0.0098	0.1682	**0.5115**	0.8085	0.9427	*0.1046*
Goat	0.0078	0.1472	**0.5023**	0.8057	0.9418	*0.1029*
Cow	0.0137	0.1519	**0.5030**	0.8063	0.9425	*0.1033*
Buffalo	0.0039	0.1396	**0.4943**	0.8046	0.9423	*0.1017*
Wolf	0.0020	0.1226	**0.4823**	0.8015	0.9418	*0.0977*
Tiger	0.0039	0.1182	**0.4767**	0.7977	0.9407	*0.0953*
Leopard	0.0020	0.1199	**0.4786**	0.8007	0.9413	*0.0971*
India_rhinoceros	0.0107	0.1599	**0.5135**	0.8089	0.9425	*0.1063*
White_rhinoceros	0.0127	0.1538	**0.5092**	0.8087	0.9425	*0.1040*
Harborseal	0.0127	0.1414	**0.4919**	0.8044	0.9420	*0.1011*
Gray_seal	0.0127	0.1375	**0.4915**	0.8040	0.9422	*0.1005*
African_elephant	0.0127	0.1553	**0.5026**	0.8045	0.9420	*0.1028*
Asiatic_elephant	0.0068	0.1448	**0.5018**	0.8047	0.9419	*0.1031*
Black_bear	0.0020	0.1106	**0.4698**	0.7979	0.9408	*0.0946*
Brown_bear	0	0.1150	**0.4702**	0.7984	0.9413	*0.0935*
Polar_bear	0.0020	0.1111	**0.4694**	0.7979	0.9412	*0.0928*
Rabbit	0.0166	0.1538	**0.4987**	0.8048	0.9418	*0.1029*
Hedgehog	0.0146	0.1780	**0.5298**	0.8157	0.9433	*0.1107*
Norway_rat	0.0156	0.1853	**0.5279**	0.8123	0.9434	*0.1106*
Vole	0.0088	0.1470	**0.5029**	0.8070	0.9427	*0.1027*
Squirrel	0.0205	0.1780	**0.5173**	0.8096	0.9426	*0.1068*
sigma	0.0048	0.0185	0.0158	0.0046	0.0007	0.0042
mu	0.0087	0.1463	0.5005	0.8061	0.9422	0.1022
sigma/mu (σ/μ)	0.5534	0.1263	**0.0316**	0.0057	0.0007	**0.0415**

For each kind of 'linkage', the cophenetic correlation coefficient can be served as stability index, through which we can measure the consensus degree between the pair-wise distances and the derived dendrogram. In general, the greater coefficient indicates that that kind of 'linkage' is better consensus. Therefore, the variation for cophenetic correlation coefficient can help us to select the optimal 'linkage' when the magnitude of the coefficient reaches the summit. These results are shown in Fig. 5.

From Fig. 5, it can be seen that the coefficient value *c* reaches the summit in the third case, which implies that 'average linkage' is the best one among these seven kinds of linkages. Thus, we can conclude that the 'average linkage' is optimal. Meanwhile, according to the result from Fig. 4, we use the '*cosine*' metric to compute the pair-wise distance between every two feature vectors via 3 segmented 7-mer. Fig. 6 illustrates the dendrogram, from which we can find that these 34 species are separated clearly:

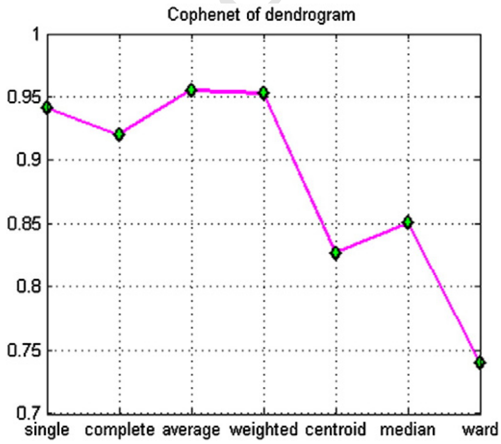

Fig. 5. The optimal linkage for dendrogram using the 3 segmented 7-mer for the mitochondrial genome sequence form 34 mammalian species.

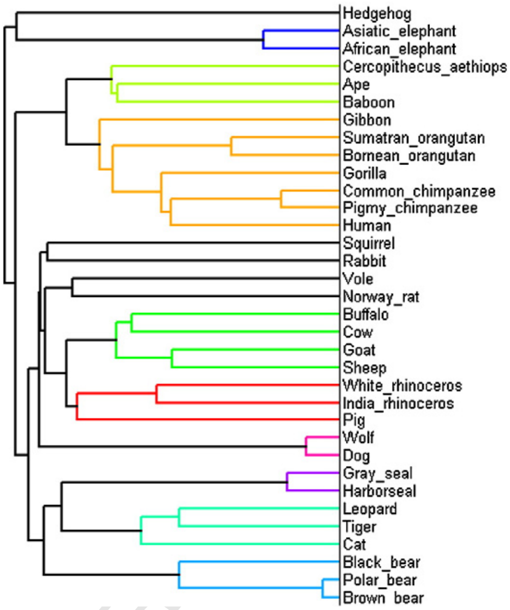

Fig. 6. The dendrogram based on 3 segmented 7-mer using 34 mitochondrial genome sequences of mammalian species.

(a) Group (Hedgehog, (Asiatic Elephant, African Elephant)) is far away from other clusters;
(b) Ten Primates are clustered closely;
(c) The Rodents (Squirrel, Vole and Rat) stand more nearer to Rabbit;
(d) The groups of Artiodactyls (Cow, Goat, and Sheep) are also close to (Rhinoceros and Pig);
(e) Dog and Wolf are close to each other, and they both belong to the Canis group;
(f) The group (Leopard, Tiger, and Cat) is clustered closely with each other;
(g) Ursidae group (Black Bear, Brown Bear, and Polar Bear) are clearly classified.

Meanwhile, these are in agreement with the evolutional facts (Cao et al., 1998; Li et al., 2001; Otu and Sayood, 2003). Therefore, it can be seen that the proposed approach is effective in comparison of genome sequences. In particular, our result suggests that the insectivore Hedgehog is the earliest species diverged out among these mammalian. This suggestion is also in accordance with those found in Krettek et al. (1995).

4. Conclusions

In this study, we investigated the optimal combination of the number of *K*-mer's order with the distance metric, i.e. $(K^*, \theta^*) =$ (7, 'euclidean') upon the data set comprising 34 mammalian mitochondrial genome sequences. Under this circumstance, we proposed an approach that improved the traditional *K*-mer, i.e. segmented *K*-mer, so as to raise the precision in similarity analysis. Then, we explored that 3 segmented 7-mer achieved the optimal results. Additionally, we found that 'average' is the optimal linkage, through which we obtained dendrogram for the genome sequence data set.

Please cite this article as: Yu, H.-J., Segmented *K*-mer and its application on similarity analysis of mitochondrial genome sequences, Gene (2013), http://dx.doi.org/10.1016/j.gene.2012.12.079

Results demonstrate that the proposed *s*-*K*-mer method outperforms the traditional *K*-mer approach. In the future, we are planning to design a new distance metric to further improve the performance on similarity analysis.

Acknowledgments

This study is supported by the National Science Foundation of China, Nos. 61133010, 31071168, 60975005, 61005010, 30900321; the special grant of China Postdoctoral Science Foundation, No. 2012T50582; the grant of China Postdoctoral Science Foundation, No. 20100480708; and the Anhui Provincial Natural Science Foundation, under 1208085MF96.

References

Akhtar, M., et al., 2007. On DNA numerical representation for period-3 based exon prediction. 5th International Workshop on Genomic Signal Processing and Statistics (Tuusula, Finland).

Bielińska-Wąż, D., 2011. Graphical and numerical representations of DNA sequences: statistical aspects of similarity. J. Math. Chem. 49 (10), 2345–2407.

Brodzik, A.K., Peters, O., 2005. Symbol-balanced quaternionic periodicity transform for latent pattern detection in DNA sequences. Proceedings of IEEE ICASSP, Philadelphia.

Cao, Y., et al., 1998. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J. Mol. Evol. 47, 307–322.

Cristea, P.D., 2003. Large scale features in DNA genomic signals. Sign. Process. 83, 871–888.

Gao, L., Qi, J., 2007. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. BMC Evol. Biol. 7, 41.

Huang, G., et al., 2011. Alignment-free comparison of genome sequences by a new numerical characterization. J. Theor. Biol. 281 (1), 107–112.

Jeffrey, H.J., 1990. Chaos game representation of gene structure. Nucleic Acids Res. 18 (8), 2163–2170.

Jun, S.R., et al., 2009. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. Proc. Natl. Acad. Sci. 107 (1), 133–138.

Karlin, S., Burge, C., 1995. Dinucleotide relative abundance extremes:a genomic signature. Trends Genet. 11, 283–290.

Krettek, A., et al., 1995. Sequence analysis of the complete mitochondrial DNA molecule of the hedgehog, Erinaceus europaeus, and the phylogenetic position of the Lipotyphla. J. Mol. Evol. 41 (6), 952–957.

Li, M., et al., 2001. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics 17 (2), 149–154.

Liao, B., Wang, T.-m, 2004a. Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation. Chem. Phys. Lett. 388 (1–3), 195–200.

Liao, B., Wang, T.M., 2004b. New 2D graphical representations of DNA sequences. J. Comput. Chem. 25, 1364–1368.

Liao, B., et al., 2005. Application of 2-D graphical representation of DNA sequence. Chem. Phys. Lett. 414, 296–300.

Otu, H.H., Sayood, K., 2003. A new sequence distance measure for phylogenetic tree construction. Bioinformatics 19 (16), 2122–2130.

Qi, J., et al., 2004. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. J. Mol. Evol. 58, 1–11.

Randić, M., 2004. Graphical representations of DNA as 2-D map. Chem. Phys. Lett. 386, 468–471.

Randić, M., 2008. Another look at the chaos-game representation of DNA. Chem. Phys. Lett. 1–3, 84–88.

Randić, M., et al., 2003a. Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem. Phys. Lett. 368, 1–6.

Randić, M., et al., 2003b. Analysis of similarity/dissimilarity of DNA sequences based on a novel 2-D graphical representation. Chem. Phys. Lett. 371, 202–207.

Randić, M., et al., 2003c. Compact 2-D graphical representation of DNA. Chem. Phys. Lett. 373 (5–6), 558–562.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4 (4), 406–425.

Sims, G.E., et al., 2009a. Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. Proc. Natl. Acad. Sci. 106 (40), 17077–17082.

Sims, G.E., et al., 2009b. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc. Natl. Acad. Sci. 106 (8), 2677–2682.

Song, J., Tang, H.W., 2005. A new 2-D graphical representation of DNA sequences and their numerical characterization. J. Biochem. Biophys. Methods 63 (3), 228–239.

Voss, R.F., 1992. Evolution of long-rang fractal correlations and 1/f noise in DNAbase sequences. Phys. Rev. Lett. 68, 3805–3808.

Wang, S., et al., 2009. Applications of representation method for DNA sequences based on symbolic dynamics. J. Mol. Struct. THEOCHEM 909, 33–42.

Wu, T.-J., et al., 1997. A measure of DNA sequence dissimilarity based on mahalanobis distance between frequencies of words. Biometrics 53 (4), 1431–1439.

Wu, T.-J., et al., 2001. Statistical measures of DNA dissimilarity under Markov chain models of base composition. Biometrics 57, 441–448.

Wu, T.-J., et al., 2005. Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences. Bioinformatics 21 (22), 4125–4132.

Xu, R., Wunsch II, Donald, 2005. Survey of clustering algorithms. IEEE Trans. Neural Netw. 16 (3), 645–678.

Yang, L., et al., 2012. Alignment free comparison: similarity distribution between the DNA primary sequences based on the shortest absent word. J. Theor. Biol. 295, 125–131.

Yao, Y.-h, Wang, T.-m, 2004. A class of new 2-D graphical representation of DNA sequences and their application. Chem. Phys. Lett. 398 (4–6), 318–323.

Yu, C., et al., 2010. A novel construction of genome space with biological geometry. DNA Res. 17 (3), 155–168.

Zhang, R., Zhang, C.T., 1994. Z curves, an intutive tool for visualizing and analyzing the DNA sequences. J. Biomol. Struct. Dyn. 11 (4), 767–782.

Please cite this article as: Yu, H.-J., Segmented *K*-mer and its application on similarity analysis of mitochondrial genome sequences, Gene (2013), http://dx.doi.org/10.1016/j.gene.2012.12.079