World Scientific Research Journal (WSRJ) ISSN: 2472-3703
www.wsr-j.org

A Novel Numerical Characterization of DNA Sequences Based on Two-Base and Its Application

Xiaolong Xue[1, a], Yunxiu Zhao[1, b], Lei Wang[1, c], Xiaoli Xie[1, d, *]

[1]College of Science, Northwest A&F University, Yangling, Shanxi, PR China

[a]741945432@qq.com, [b]2903060323@qq.com, [c]983222671@qq.com,

[d]xiemary@nwsuaf.edu.cn

Abstract: Analyzing DNA sequences is a topic in bioinformatics. Traditionally, comparing DNA sequence is carried out by alignment method. However, it is extremely complex in time and space complexity. In the paper, a novel alignment-free method is proposed based on the position information of two adjacent nucleotides. A DNA sequence is transformed to a 48D vector, which includes frequency, mean value and variance of position for each two bases. The Euclidean distances for new vectors are calculated to carry on the similarity analysis. Finally, comparing Clustal W method with double nucleotides vector and single nucleotide vector.

Keywords: Numerical characterization, DNA sequence, two-base, cluster analysis, Similarity analysis.

---

1. INTRODUCTION

The similarity analysis of biologic sequences is a hot topic in recent years. It plays an important role on being aware of the evolutionary relationship between DNA sequences. A great deal of methods have been successfully applied to classify the sequences into certain types [1-6]. For the moment, the main researches are divided into sequence alignment method and sequence alignment-free method. Nevertheless, alignment method is extremely complex in time and space complexity although it can guarantee the accuracy of classifying. The k-mer method is very popular of alignment-free methods. However, it only considers the frequency of the k-word, so it leads to the loss of information of DNA sequences [7-8]. In addition, other alignment-free method can be categorized into several classes in general: (1) Methods based on substrings carry out the similarity in a pair of sequences [9]. (2) Alignment-free sequence analysis and comparison can be successfully made according to information theory. Existing information theory include global and local characterization of DNA, estimating genome entropy to motif and region classification [10]. (3) Graphical approaches are extremely useful in dealing with various biological problems, especially for very complicated biological systems due to intuitive insights. (4) Some properties are integrated into sequence

alignment-free method [11-14]. Composition vectors based on k-word position is a new method. Many researchers have begun to extract the position information of a k-word [15]. His lily and Zhao Xin proposed an alignment-free method based on position of each nucleotide and amino acid. For mining the more information, we present a new numerical characterization based on the position information of two adjacent nucleotides.

In the paper, we build a new 48 dimensional numerical vector to character a DNA protein. We take into account the positions and frequency of two adjacent nucleotides. To test the efficiency of our method, our method and alignment method (ClustalW) are compared.

2. MATERIALS AND METHOD

2.1 Numerical characterization

Let $S=(s_1,s_2,\cdots,s_N)$ be a DNA sequence, $s_i \in \{A,T,C,G\}$. In the paper, a DNA sequence is transformed to a binary indicator sequence by coding adjacent nucleotides, we define W_{AA} as follows

$$W_{AA}(i)=\begin{cases}1 & \text{if } AA \text{ is present at location i of the sequence}\\ 0 & \text{otherwise}\end{cases}$$

W_{AC}, W_{AG}, W_{AT} et al. are defined similarly.

If a sequence is AACGTAGTCAA, the corresponding indicator sequence of nucleotide AA is $W_{AA}=10000000010$. It is worth mentioning that the last nucleotide and the first nucleotide doesn't form a circle.

For obtaining more information of DNA sequence, we construct three characterizations f_k, μ_k, D_k ($k=AA,AT,AC,AG,\cdots,GG$), and they describe the frequency, the average position and variation of position for each two-base.

$$f_{AA}=\frac{\sum_{i=1}^{N}W_{AA}(i)}{N} \qquad \mu_{AA}=\frac{\sum_{i=1}^{N}iW_{AA}(i)}{N} \qquad D_{AA}=\frac{1}{N-1}\sum_{i=1}^{N}(iW_{AA}(i)-\mu_{AA})^2$$

These feathers form a 48 dimension vector:

$$L=\left(f_{AA},\mu_{AA},D_{AA},f_{AC},\mu_{AC},D_{AC},\cdots,f_{TT},\mu_{TT},D_{TT}\right).$$

In order to avoid the influence of extremum, vector L is standardized. Take the first component f_{AA} as an example:

If $f_{\min}\neq f_{\max}$, $f'=\frac{f-f_{\min}}{f_{\max}-f_{\min}}$;

If $f_{\min}=f_{\max}$, $f'=f_{\min}$.

Where, $f_{\min}$, $f_{\max}$ is the minimum and maximum of f_{AA}, respectively. f' Is the first component which is normalized? Therefore, a novel vector who's every component's value ranges from 0 to 1 can be gained.

2.2 Clustering analysis

The similarity matrix can be obtained by computing the Euclidean distances between any two vectors. On this basis of similarity matrix, the phylogenetic trees can be constructed.

2.3 Data sets

The new alignment-free method based on two-base is tested on different data sets such as 30 gene sequence of mammal and 48 hepatitis E virus (HEV) sequence from GenBank.

3. RESULTS

3.1 30 mammals sequences

The new method is first tested on a mitochondrial DNA data set of 30 mammalian genomes, and each sequence has a length range from 16,300 to 17,500 nucleotides. Fig.1 shows the clustering result.

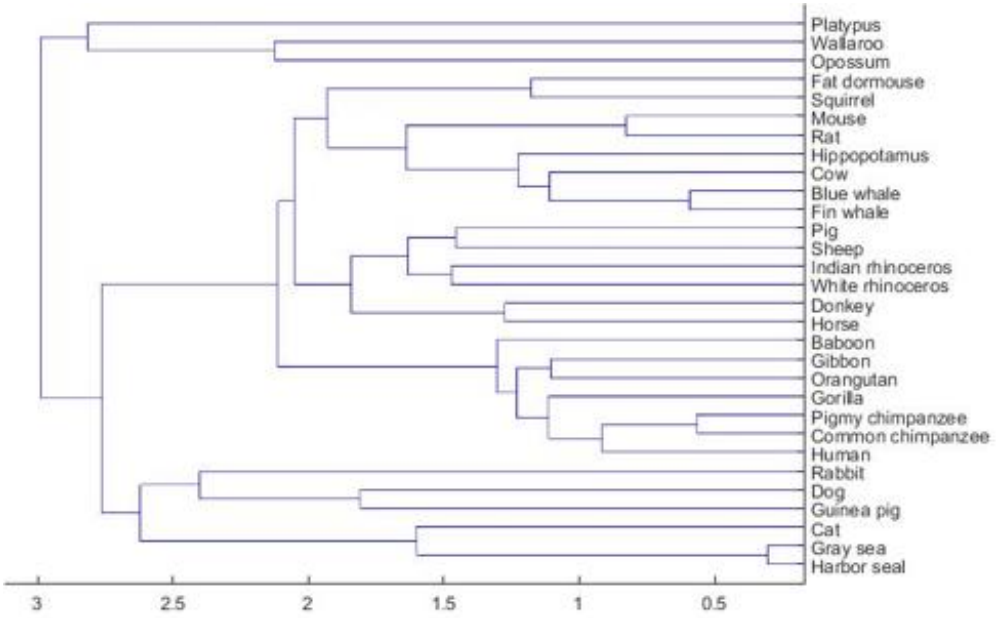

Fig 1. Phylogenetic tree of 30 mammalian genomes by two-base method

Fig.1 shows the classification results of 30 mammalian genomes in detail. For example, human, common chimpanzee and pigmy chimpanzee belong to a branch, and it matches the basic biological information.

3.2 48 hepatitis E virus (HEV) sequence

The research about virus not only benefits personal health in a deep degree, but also concerns the entire ecosystem in a sense. Hepatitis E virus which easily lead to hepatitis are the most dangerous for modern life style. Expressly two subtypes will be chosen to test the efficiency of our method. The result of phylogenetic is drawn as follows.

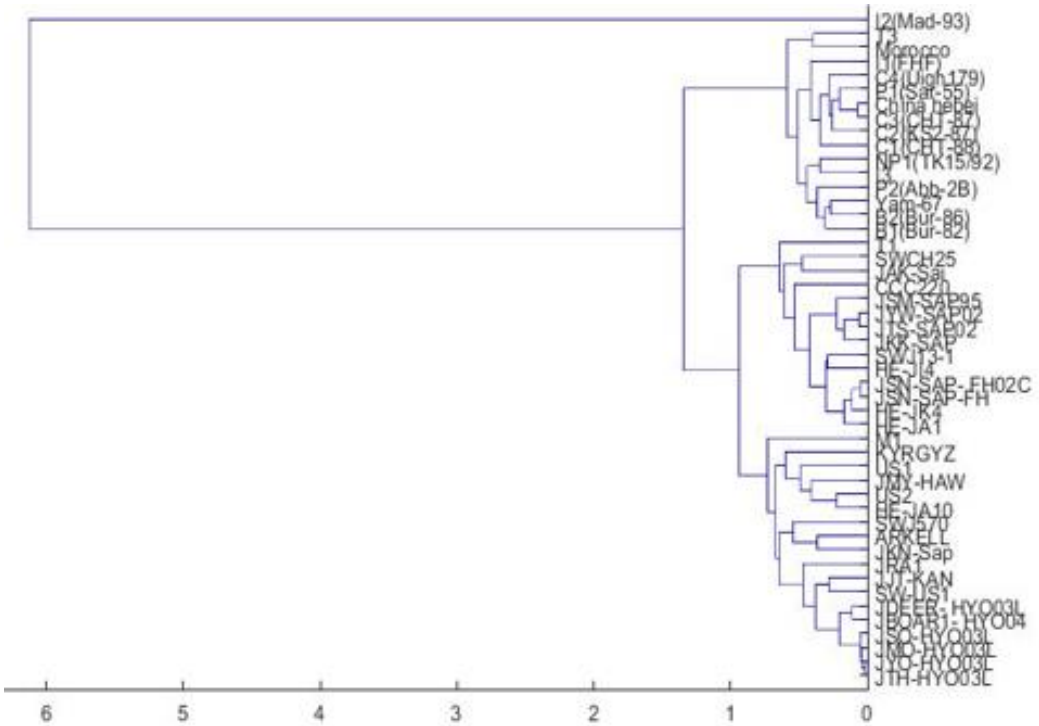

Fig 2. Phylogenetic tree of 48 hepatitis E virus (HEV) genomes by two-base method

Fig.2 shows the clustering results of 48 hepatitis E virus (HEV) genomes. It is obvious that the results have a strong similarity because two-base method can extract more information. The first hepatitis E virus (HEV) genome at the top of the figure is separated, so we can think it belong to a specific class. Similarly, the above results can offer some properties we need.

3.3 Method comparison

To further confirm the proposed method is effective, the similarity distances are got by computing the Clustal W method, the two-base method and the single-base method for the same data sets. The correlation coefficients are displayed in the below table (take mammalian genome sequences as an example):

4. DISCUSSION

Position information is one of the most important information hidden in the original sequences. By getting more messages, the proposed method based on two-base is more effective than single-base. The results of clustering analysis according to the new vector indicate ergodic base can offer a new thinking to analyze DNA sequences and protein sequences. The vector characterization involved in the research [16-18] for protein sequences can draw more information to analyze corresponding sequences.

It is no doubt that more information of original sequence will be important to make clustering analysis and similarity analysis. The new method based on two-base not only mining information of DNA sequences in a deep degree, but also obtaining easily. Consider three continuous nucleotides, the dimension of new vector will be 192 that means more complicated

data processing. Thus, the proposed method based on two adjacent nucleotide is more efficient to analyze biological sequences.

Table 1. The correlation coefficients between ClustalW & double and ClustalW & single

Species Name	ClustalW & double	ClustalW & single
Human	-0.839	-0.757
Common chimpanzee	-0.806	-0.733
Pigmy chimpanzee	-0.839	-0.755
Gorilla	-0.777	-0.709
Orangutan	-0.834	-0.794
Gibbon	-0.86	-0.813
Baboon	-0.86	-0.825
Horse	-0.646	-0.567
White rhinoceros	-0.665	-0.555
Harbor sea	-0.866	-0.695
Gray seal	-0.879	-0.697
Cat	-0.499	-0.399
Fin whale	-0.513	-0.436
Blue whale	-0.521	-0.449
Cow	-0.36	-0.34
Rat	-0.405	-0.316
Mouse	-0.493	-0.411
Opossum	-0.72	-0.519
Wallaroo	-0.46	-0.091
Platypus	-0.21	0.072
Squirrel	-0.426	-0.357
Fat dormouse	-0.383	-0.403
Guinea pig	-0.367	-0.147
Donkey	-0.521	-0.506
Indian rhinoceros	-0.701	-0.582
Dog	-0.563	-0.489
Sheep	-0.642	-0.527
Pig	-0.671	-0.629
Hippopotamus	-0.425	-0.392
Rabbit	-0.077	0.208

It is obvious that the absolute values listed in the first column are higher than the second column. In other words, the new method based on the adjacent nucleotides is more effective than single-base method. Specially, the two-base method can extract more information for long sequences.

ACKNOWLEDGEMENTS

The paper was supported by National Natural Science Foundation of China (31572361).

REFERENCES

[1] Mendizabal-Ruiz Gerardo, Román-God ínez Israel, Torres-Ramos Sulema, Salido-Ruiz Ricardo A, Velez-Perez Hugo. Genomic signal processing for DNA sequence clustering. PeerJ. 2018 Vol. Jan 24; 6: e4264.

[2] Guo-Sen Xie, Xiao-Bo Jin, Chunlei Yang, Jiexin Pu, Zhongxi Mo. Graphical representation and similarity analysis of DNA sequences based on trigonometric functions. Springer Journal. 2018, Vol. 66(2): p113-133.

[3] Xin Jin, Qian Jiang, Yanyan Chen, Shin-Jye Lee, Rencan Nie. Similarity/dissimilarity calculation methods of DNA sequences: A survey. Journal of Molecular Graphics and Modelling. 2017, Vol, 76: p 342-355.

[4] Tung Hoang, Changchuan Yin, Stephen S.-T. Yau. Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison. Genomics. 2016, Vol, 108(3-4): p134-142.

[5] Liwei Liu, Chao Li, Fenglan Bai, Qi Zhao. An optimization approach and its application to compare DNA sequences. Journal of Molecular Structure. 2015, Vol, 1082: p49-55.

[6] Fenglan Bai, Jihong Zhang, Junsheng Zheng, Chao Li, Liwei Liu. Vector representation and its application of DNA sequences based on nucleotide triplet codons. Journal of Molecular Graphics and Modelling. 2015, Vol, 62: p150-156.

[7] Chun Li, Yan Yang, Meiduo Jia, Yingying Zhang, Xiaoqing Yu. Phylogenetic analysis of DNA sequences based on k -word and rough set theory. Physica A: Statistical Mechanics and its Applications. 2014, Vol, 398: p162-171.

[8] Shuyan Ding, Yang Li, Xiwu Yang, Tianming Wang. A simple k -word interval method for phylogenetic analysis of DNA sequences. Journal of Theoretical Biology. 2013, Vol, 317: p192-199.

[9] Domazet-Lošo Mirjana, Haubold Bernhard. Alignment-free detection of local similarity among viral and bacterial genomes. Bioinformatics. 2011, Vol, 27(11): p1466-1472.

[10] Susana Vinga. Information theory applications for biological sequence analysis. Briefings in Bioinformatics. 2014, Vol, 15(3): p376-389.

[11] Tung Hoang, Changchuan Yin, Hui Zheng, Chenglong Yu, Rong Lucy He. A new method to cluster DNA sequences using Fourier power spectrum. Journal of Theoretical Biology. 2015, Vol, 372: p135-145.

[12] Mo Deng, Chenglong Yu, Qian Liang, He Rong L, Stephen S.-T. Yau. A novel method of characterizing genetic sequences: genome space with biological distance and applications. PL o S One. 2011, Vol, 6(3): e17293.

[13] Changchuan Yin, Stephen S.-T. Yau. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. Journal of Theoretical Biology. 2007, Vol, 247(4): p687-694

[14] Wei Chen, Tian-Yu Lei, Dian-Chuan Jin, Hao Lin, Kuo-Chen Chou. PseKNC: A flexible web server for generating pseudo K-tuple nucleotide composition. Analytical Biochemistry. 2014, Vol, 456: p53-60

[15] Xiwu Yang, Tianming Wang. A novel statistical measure for sequence comparison on the basis of k -word counts. Journal of Theoretical Biology. 2013, Vol, 318: p91-100

[16] Chuanyan Wu, Rui GAO, Yang De Marinis, Yusen Zhang. A novel model for protein sequence similarity analysis based on spectral radius. Journal of Theoretical Biology. 2018,

Vol, 446: p61-70

[17] Mehrotra Prachi, Ami Vimla Kany G, Srinivasan Narayanaswamy. Clustering of multi-domain protein sequences. Proteins. 2018, Vol, 86(7): p759-776

[18] Lily He, Yongkun Li, Rong Lucy He, Stephen S.-T. Yau. A novel alignment-free vector method to cluster protein sequences. Journal of Theoretical Biology. 2017, Vol, 427: p41-52.