Genomics 108 (2016) 134–142

Contents lists available at ScienceDirect

Genomics

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

Methods Paper

Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison

Tung Hoang [a], Changchuan Yin [a], Stephen S.-T. Yau [b],*

[a] Department of Mathematics, Statistics and Computer Science, University of Ilinois at Chicago, Chicago, IL 60607, USA
[b] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 7 March 2016
Received in revised form 4 August 2016
Accepted 12 August 2016
Available online 15 August 2016

Keywords:
Chaos game representation
Digital signal processing
Clustal Omega

ABSTRACT

Numerical encoding plays an important role in DNA sequence analysis via computational methods, in which numerical values are associated with corresponding symbolic characters. After numerical representation, digital signal processing methods can be exploited to analyze DNA sequences. To reflect the biological properties of the original sequence, it is vital that the representation is one-to-one. Chaos Game Representation (CGR) is an iterative mapping technique that assigns each nucleotide in a DNA sequence to a respective position on the plane that allows the depiction of the DNA sequence in the form of image. Using CGR, a biological sequence can be transformed one-to-one to a numerical sequence that preserves the main features of the original sequence. In this research, we propose to encode DNA sequences by considering 2D CGR coordinates as complex numbers, and apply digital signal processing methods to analyze their evolutionary relationship. Computational experiments indicate that this approach gives comparable results to the state-of-the-art multiple sequence alignment method, Clustal Omega, and is significantly faster. The MATLAB code for our method can be accessed from: www.mathworks.com/matlabcentral/fileexchange/57152

1. Introduction

Numerical and graphical representations have gained more and more attention in bioinformatics, especially when digital signal processing methods, such as discrete Fourier transform (DFT), are becoming increasingly important in handling large data. Since biological sequences are strings of symbolic characters, numerical values must be assigned to those sequences in order to apply techniques like DFT [1]. For example, Voss [2] proposed the so-called Voss binary indicator to convert a DNA sequence to four sequences of 0 and 1 that represent A, C, G, T correspondingly. This technique has been broadly used in recent research [3–5]. Anastassiou [6] presented complex conjugate pairs $t = a^*$ and $g = c^*$ for numerical representation. One such assignment is the following: $a = 1 + i$, $t = 1 - i$, $c = -1 - i$, $g = -1 + i$. Via this assignment, the complementary DNA strand is represented in a nice way such that all 'palindromes' will yield conjugate symmetric numerical sequences that have interesting mathematical properties. Yau et al. [7] discovered a method to represent DNA sequence without degeneracy. Using a two-dimensional graphical representation, this method provides a simple way to view, sort, and compare various gene structures.

While being used in many applications, most of the numerical representation methods are nucleotide mapping, where each nucleotide is assigned a fixed numerical value. Consequently, they all suffer from the same drawback: not being one-to-one [4]. Thus, having a representation that is sequence mapping would be better, as history of the previous part of the sequence can be stored and maintained.

Chaos Game Representation (CGR) was proposed by Jeffrey [8] as a two-dimensional graphical representation of DNA sequences using iterated function systems. Via a technique from chaotic dynamics, both global and local patterns and features of DNA sequence can be visually exhibited using CGR. Interestingly, given a CGR point on the plane, we can trace back to the origin of the sequence, thus the original DNA sequence can always be reconstructed. That means CGR is not only a nucleotide mapping but also is a sequence mapping, and it can be proved to be a true one-to-one representation. One of the most remarkable advantages of CGR is that image obtained from parts of a genome present the same structure as that of the whole genome. As a consequence, partial sequence can display genome signatures of the whole sequence, thus gives rise to possibility of comparing genomic sequences when only parts of the genomes are available [9]. Motivated by these unique properties, several researches based on CGR have been proposed recently. Almeida et al. [10] used CGR as an alignment-free comparisons of sequences based on oligonucleotide frequencies. Joseph and Sasikumar [11] used CGR for alignment-based comparisons of DNA sequences, in which a fast algorithm for identifying all local alignments

* Corresponding author.
E-mail address: yau@uic.edu (S.S.-T. Yau).

http://dx.doi.org/10.1016/j.ygeno.2016.08.002
0888-7543/© 2016 Elsevier Inc. All rights reserved.

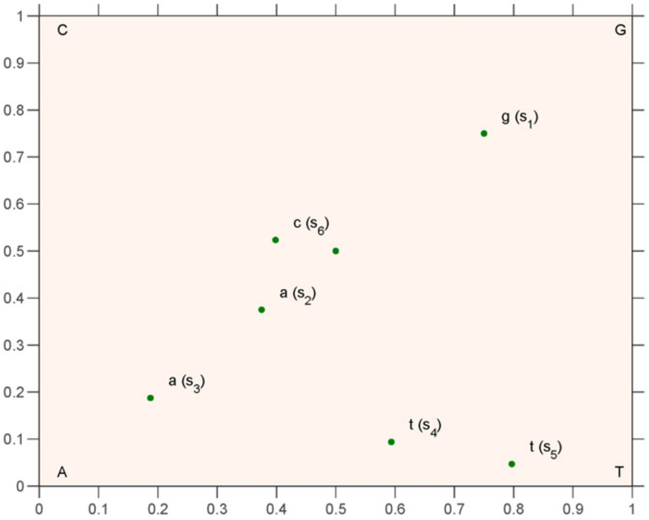

Fig. 1. CGR of $s_1s_2s_3s_4s_5s_6$ = “gaattc”.

between two DNA sequences was presented. Tanchotsrinon et al. [12] proposed two new feature extraction methods, namely ChaosCentroid and ChaosFrequency, for predicting Human Papillomavirus (HPV) genotypes associated with cervical cancer. The special features of CGR show that it would be a promising candidate for DNA representation.

DNA sequence comparison is very important in understanding evolutionary relationship in bioinformatics research. Multiple sequence alignment methods have been widely used to classify genes and proteins [13,14]. These pioneering approaches often rely on optimal alignments by using selected scoring systems via a substitution matrix and some evolutionary probability of mutation [15]. Even though these methods can give accurate classification, generally the running time is significantly high due to their computational complexity. Therefore, alignment-free approach has been extensively studied during recent years [16–18]. Unlike multiple sequence alignment, alignment-free approach is less dependent on evolutionary models, and does not assume that homologous regions are contiguous. Consequently, it is computationally inexpensive and more applicable. For example, in k-mer method [16,19], the set of k-mers, or subsequences of length k, in each DNA sequence are collected; and the evolutionary distance between DNA sequences are computed based on the corresponding sets of k-mers. With the choice of some appropriate value of k, k-mer method can give a fairly good biological classification with reasonable running time [16,20,21].

Discrete Fourier transform (DFT) is one of the most common used tools in digital signal processing. Via the use of DFT, a sequence in time domain is converted into a new one in frequency domain in such a way that hidden properties can be revealed. With this characteristic, DFT has been used in numerous DNA researches, such as gene prediction [22], protein coding region [23], and periodicity analysis [24].

Given the above advantages of Chaos Game Representation and DFT, it would be beneficial to consider CGR as a numerical representation in DFT in DNA research. In this paper, we employ CGR as a complex number representation to construct a new alignment-free method to classify DNA sequences based on DFT power spectrum. UPGMA (Unweighted Pair Group Method with Arithmetic Mean, Sokal [25]) clustering method is then performed on power spectra to determine the evolutionary relationship between DNA sequences. The method is tested and compared to the state-of-the-art multiple sequence alignment method Clustal Omega [26] on various datasets for speed and accuracy.

2. Materials and methods

2.1. Chaos game representation

Chaos Game Representation is able to give both numerical and graphical representation for genomic sequences [8]. Given the unit square in the Euclidean plane, the binary CGR vertices were assigned to the four nucleotides as A = (0,0); C = (0,1); G = (1,1); T = (1,0). Starting with the center of the square, $(\frac{1}{2}, \frac{1}{2})$, the CGR position of each nucleotide of the DNA sequence is calculated by moving a pointer to half the distance between the previous point and the corner square of the current nucleotide. An illustration for the CGR of a short sequence, namely “gaattc”, is provided in Fig. 1. The first nucleotide $s_1 = g$ is the

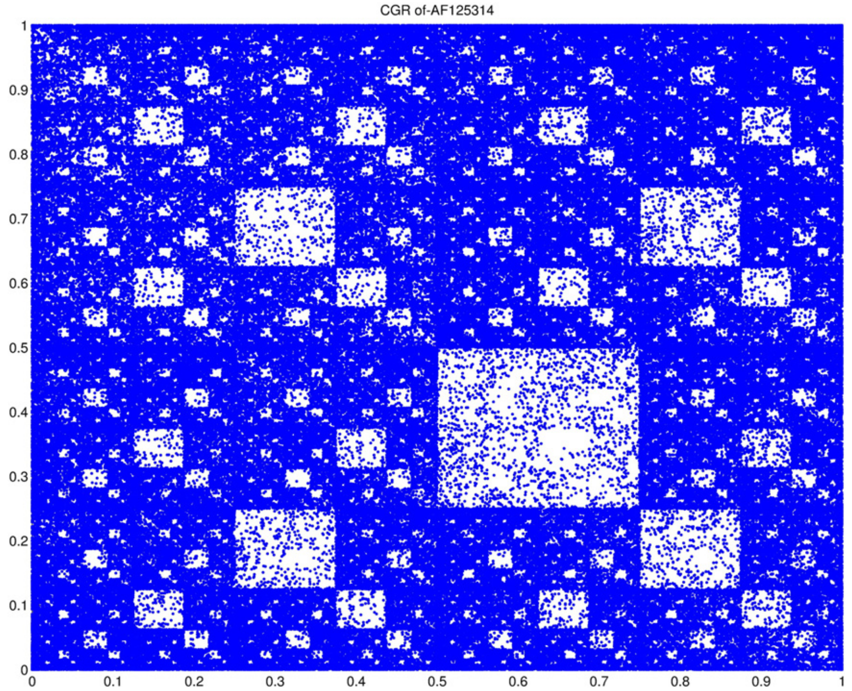

Fig. 2. CGR of *Mus musculus* chromosome X, GenBank ID: AF125314.

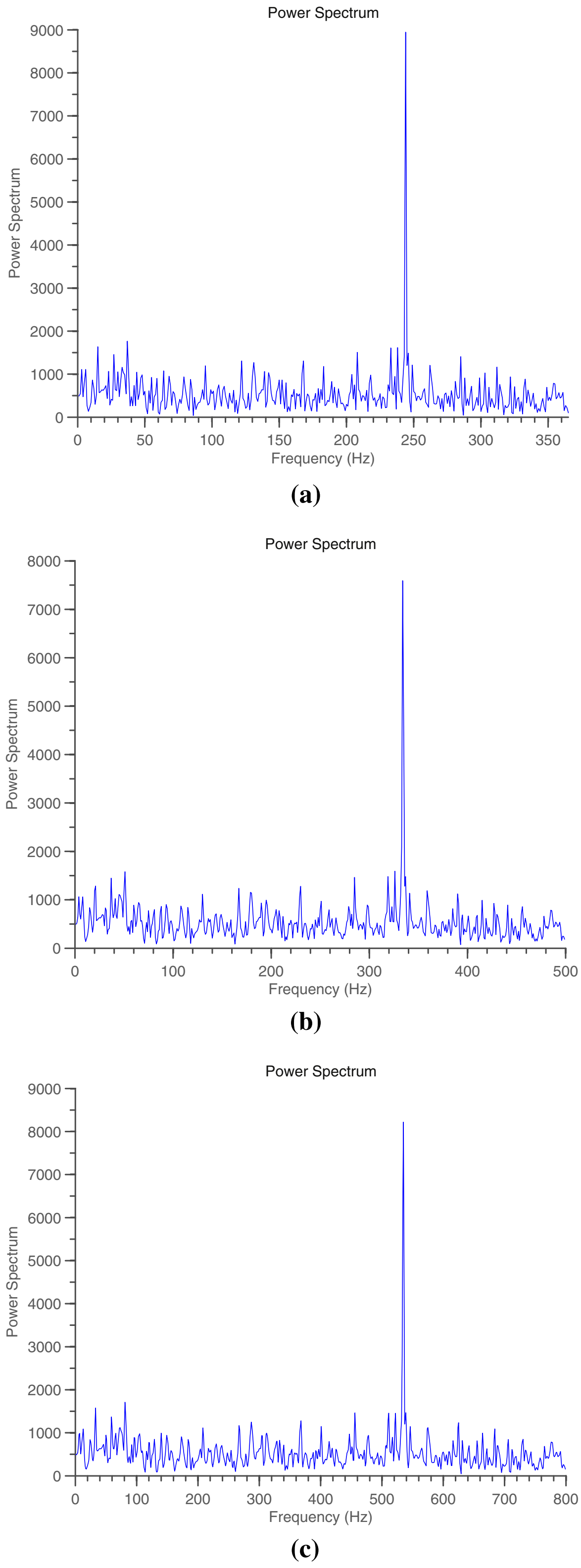

Fig. 3. DFT Power Spectrum of *Bubo bubo* voucher NHMO-BC120 cytochrome oxidase subuinit 1 (COI) gene (GenBank ID: GU571285): original sequence (a); even scaled to 1000 bp (b); even scaled to 1600 bp (c). Only first half of the spectrum is plotted due to symmetry.

midpoint of the segment made from the center and the G corner of the square. The second nucleotide $s_2 = a$ is the midpoint of s_1 and the A corner of the square. The third nucleotide $s_3 = a$ is the midpoint of s_2 and the A corner of the square, and so on.

The formal definition of CGR is given by an iterated function as in the equations below. For a DNA sequence $s_1s_2 \ldots s_n \ldots$, the corresponding CGR sequence $(X_n) = (x_n, y_n)$ is given by:

$$X_0 = \left(\frac{1}{2}, \frac{1}{2}\right), X_n = \frac{1}{2}(X_{n-1} + W)$$

where W is coordinates of the corners of the unit square $A = (0,0); C = (0,1); G = (1,1); T = (1,0)$ if s_n is a, c, g, t respectively.

The unit square is divided into four quadrants such that the lower and upper halves indicate the base composition, and the diagonals indicate the purine/pyrimidine composition [9]. By definition, there is a one-to-one correspondence between subsequences counted from the start of a DNA sequence and points of the CGR. Therefore, up to the resolution of the screen, each point in the CGR corresponds to exactly one subsequence that starts from the first base. Thus, the source sequence can be recovered from the coordinates. The above argument is summarized in the following theorem. We provide the proof of the theorem in supplementary document.

Theorem 2.1. Given an exact coordinate of a CGR point on the plane, we can determine its corresponding nucleotide, as well as its respective position in the original DNA sequence. Thus with just this one point, the whole original nucleotide subsequence up to the current nucleotide can be reconstructed.

By the unique properties of CGR, subsequences of a gene or genome exhibit the main characteristics of the whole sequence, thus it is useful in detection of special genome features, especially when parts of genome have not been available yet [9]. In many cases, obvious pattern can be observed in the representation of a DNA sequence. Therefore, it would be beneficial for researchers to discover interesting properties of the sequence, which in turn might lead to meaningful conclusion that cannot be detected by computers.

As an example of 2-D fractal pattern of CGR of a DNA sequence, the CGR of *Mus musculus* chromosome X (GenBank ID: AF125314) is shown in Fig. 2. The largest white square indicates the scarcity of CT in the DNA sequence. The second to largest four white squares indicate the consequence of the lack of CT, i.e. the scarcity of CTA, CTC, CTG, CTT. Similarly, the third to largest white squares would indicate the scarcity of CTXX in the DNA sequence, where $X \in \{A, C, G, T\}$.

2.2. DFT and power spectrum

In signal processing, sequences in time domain are commonly transformed into frequency domain to make some important features visible. Via that one-to-one transformation, no information is lost but some hidden properties could be revealed. One of the most common transformation is discrete Fourier transform [27]. For a signal of length N, $f(n)$, $n = 0, \ldots, N-1$, the DFT of the signal at frequency k is

$$F(k) = \sum_{n=0}^{N-1} f(n) e^{-i\frac{2\pi}{N}kn}$$

for $k = 0, \ldots, N-1$.

The DFT power spectrum of a signal at frequency k is defined as

$$PS(k) = |F(k)|^2, k = 0, \ldots, N-1$$

Notice that by definition, $PS(0) = |F(0)|^2 = |\sum_{n=0}^{N-1} f(n)|^2$.

The DFT is often used to find the frequency components of a signal buried in a noisy time domain. For example, below is the DFT Power Spectrum of *Bubo bubo* voucher NHMO-BC120 cytochrome oxidase subunit 1 (COI) gene (GenBank ID: GU571285), with length $N = 720$ bp (Fig. 3a). As it appears on the graph, there is a peak at frequency $N/3$

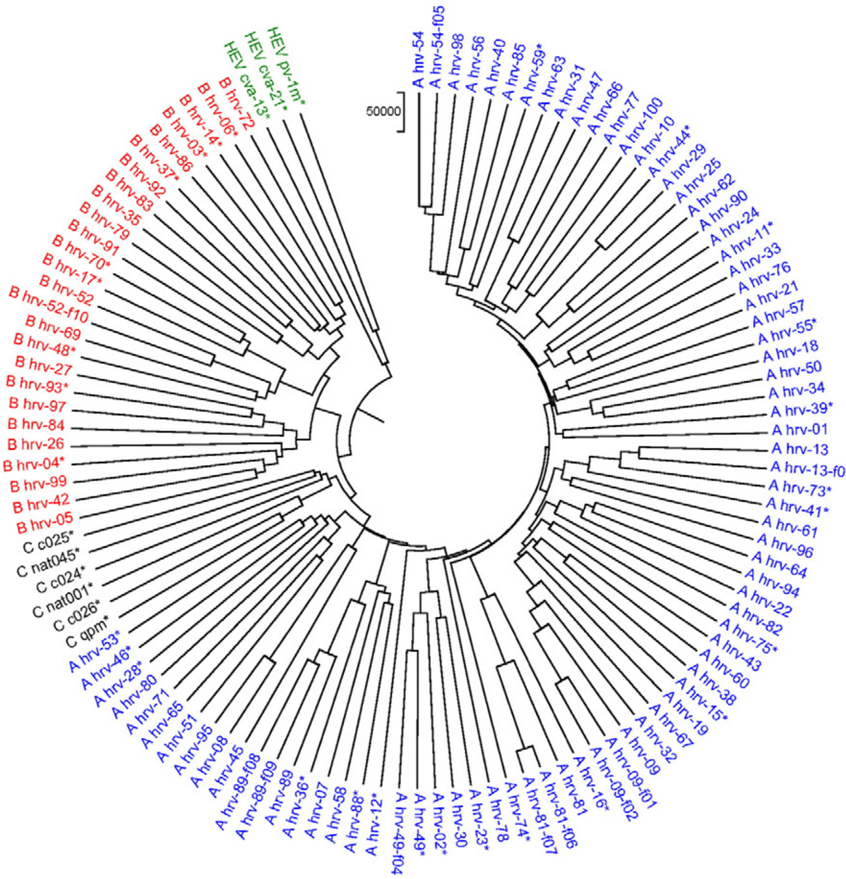

Fig. 4. Phylogenetic tree of 113 HRV genomes and 3 outgroup sequences based on our method.

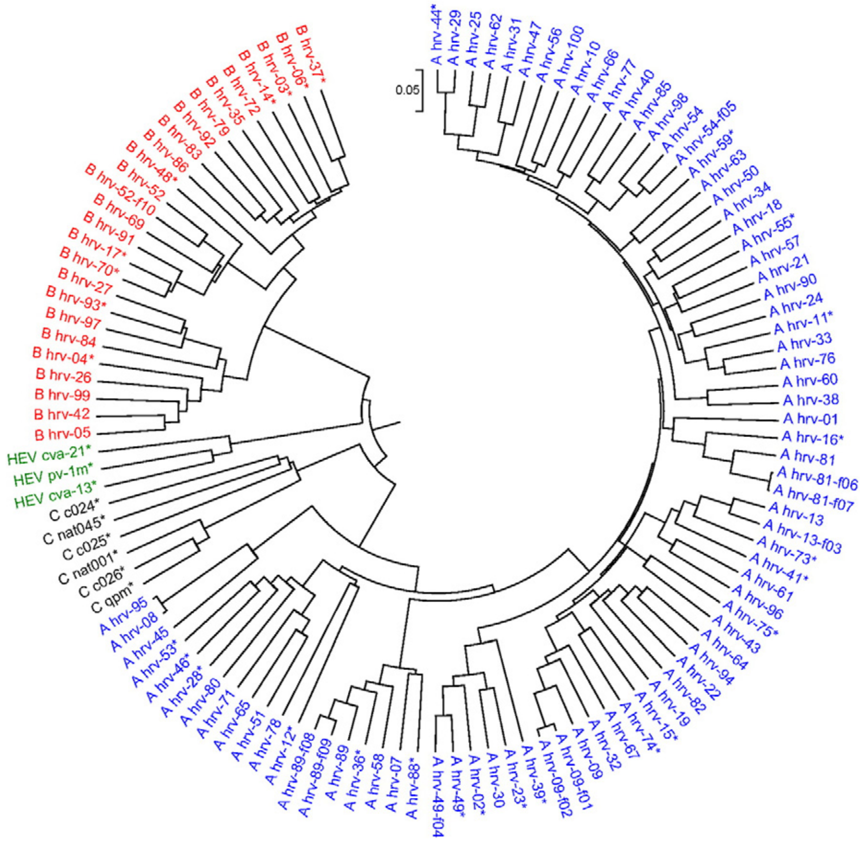

Fig. 5. Phylogenetic tree of 113 HRV genomes and 3 outgroup sequences based on Clustal Omega.

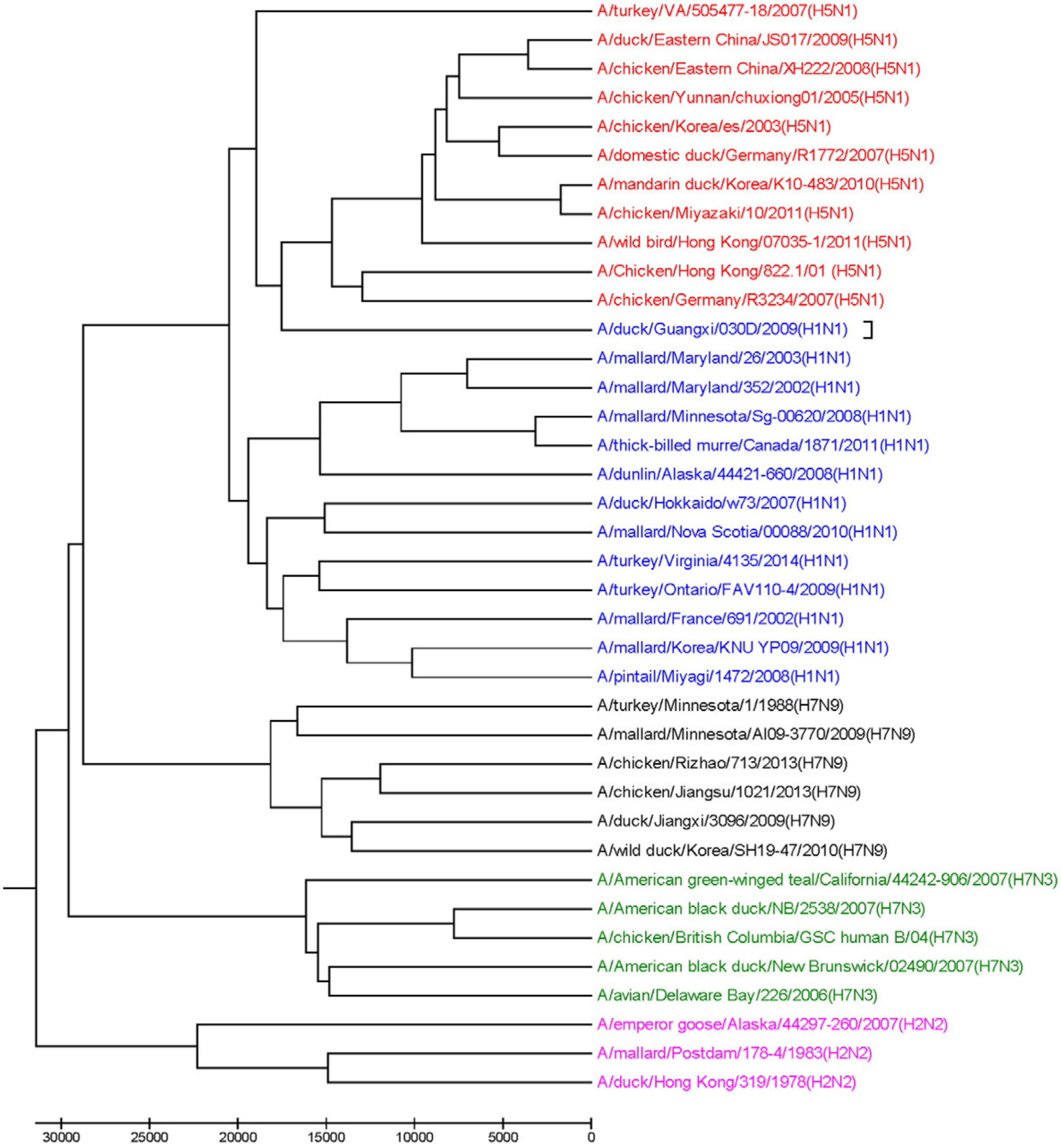
A/turkey/VA/505477-18/2007(H5N1)
A/duck/Eastern China/JS017/2009(H5N1)
A/chicken/Eastern China/XH222/2008(H5N1)
A/chicken/Yunnan/chuxiong01/2005(H5N1)
A/chicken/Korea/es/2003(H5N1)
A/domestic duck/Germany/R1772/2007(H5N1)
A/mandarin duck/Korea/K10-483/2010(H5N1)
A/chicken/Miyazaki/10/2011(H5N1)
A/wild bird/Hong Kong/07035-1/2011(H5N1)
A/Chicken/Hong Kong/822.1/01 (H5N1)
A/chicken/Germany/R3234/2007(H5N1)
A/duck/Guangxi/030D/2009(H1N1)
A/mallard/Maryland/26/2003(H1N1)
A/mallard/Maryland/352/2002(H1N1)
A/mallard/Minnesota/Sg-00620/2008(H1N1)
A/thick-billed murre/Canada/1871/2011(H1N1)
A/dunlin/Alaska/44421-660/2008(H1N1)
A/duck/Hokkaido/w73/2007(H1N1)
A/mallard/Nova Scotia/00088/2010(H1N1)
A/turkey/Virginia/4135/2014(H1N1)
A/turkey/Ontario/FAV110-4/2009(H1N1)
A/mallard/France/691/2002(H1N1)
A/mallard/Korea/KNU YP09/2009(H1N1)
A/pintail/Miyagi/1472/2008(H1N1)
A/turkey/Minnesota/1/1988(H7N9)
A/mallard/Minnesota/AI09-3770/2009(H7N9)
A/chicken/Rizhao/713/2013(H7N9)
A/chicken/Jiangsu/1021/2013(H7N9)
A/duck/Jiangxi/3096/2009(H7N9)
A/wild duck/Korea/SH19-47/2010(H7N9)
A/American green-winged teal/California/44242-906/2007(H7N3)
A/American black duck/NB/2538/2007(H7N3)
A/chicken/British Columbia/GSC human B/04(H7N3)
A/American black duck/New Brunswick/02490/2007(H7N3)
A/avian/Delaware Bay/226/2006(H7N3)
A/emperor goose/Alaska/44297-260/2007(H2N2)
A/mallard/Postdam/178-4/1983(H2N2)
A/duck/Hong Kong/319/1978(H2N2)
30000
25000
20000
15000
10000
5000
0

Fig. 6. Phylogenetic tree of segment 6 of Influenza A virus genomes based on our method.

due to the 3-base periodicity property of DNA sequence, which can be detected only in frequency domain.

2.3. Even scaling method

Power spectrum of a DNA sequence has the same length as in the original sequence. Thus, it is difficult to make a direct comparison of two power spectra of DNA sequences. A good way to resolve this problem is to use even scaling method by Yin and Yau [4], where a data series can be stretched to a longer one in a way that still preserves the property of the original series. Specifically, if T_n is the power spectrum of a DNA sequence of length n, and m is some integer such that $m>n$, then T_m is defined as follows:

$$T_m(k) = \begin{cases} T_n(Q) & \text{if } Q \in Z^+ \\ T_n(R) + (Q-R)(T_n(R+1)-T_n(R)) & \text{if } Q \notin Z^+, \\ \text{where } Q = \dfrac{kn}{m}, R = \left\lfloor \dfrac{kn}{m} \right\rfloor & \end{cases}$$

With this even scaling method, DNA sequences of different lengths can be compared easily, e.g. all sequences could be scaled to the maximum length of the underlying data set. In other words, DNA sequences can be embedded into the same Euclidean space. The illustration of even scaling method on the above DFT Power Spectrum of *Bubo bubo* voucher NHMO-BC120 cytochrome oxidase subuinit 1 (COI) gene is given in Fig. 3.

2.4. Method

The detailed method to give the evolutionary tree of DNA sequences is described in the algorithm below.

2.4.1. Input

A set of n DNA sequences.

2.4.2. Output

Phylogenetic tree representing evolutionary relations between those n DNA sequences.

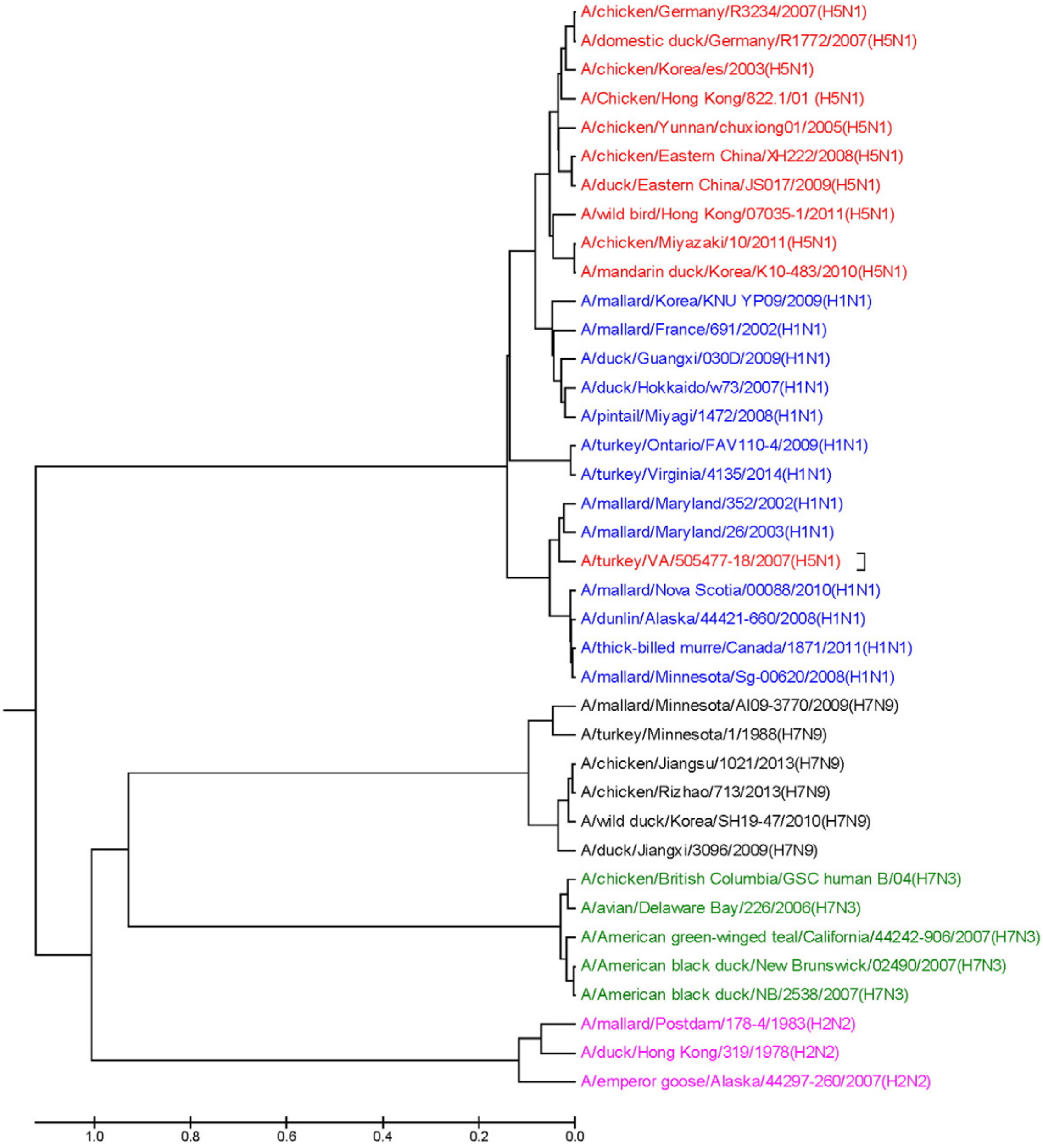
A/chicken/Germany/R3234/2007(H5N1)
A/domestic duck/Germany/R1772/2007(H5N1)
A/chicken/Korea/es/2003(H5N1)
A/Chicken/Hong Kong/822.1/01 (H5N1)
A/chicken/Yunnan/chuxiong01/2005(H5N1)
A/chicken/Eastern China/XH222/2008(H5N1)
A/duck/Eastern China/JS017/2009(H5N1)
A/wild bird/Hong Kong/07035-1/2011(H5N1)
A/chicken/Miyazaki/10/2011(H5N1)
A/mandarin duck/Korea/K10-483/2010(H5N1)
A/mallard/Korea/KNU YP09/2009(H1N1)
A/mallard/France/691/2002(H1N1)
A/duck/Guangxi/030D/2009(H1N1)
A/duck/Hokkaido/w73/2007(H1N1)
A/pintail/Miyagi/1472/2008(H1N1)
A/turkey/Ontario/FAV110-4/2009(H1N1)
A/turkey/Virginia/4135/2014(H1N1)
A/mallard/Maryland/352/2002(H1N1)
A/mallard/Maryland/26/2003(H1N1)
A/turkey/VA/505477-18/2007(H5N1)
A/mallard/Nova Scotia/00088/2010(H1N1)
A/dunlin/Alaska/44421-660/2008(H1N1)
A/thick-billed murre/Canada/1871/2011(H1N1)
A/mallard/Minnesota/Sg-00620/2008(H1N1)
A/mallard/Minnesota/AI09-3770/2009(H7N9)
A/turkey/Minnesota/1/1988(H7N9)
A/chicken/Jiangsu/1021/2013(H7N9)
A/chicken/Rizhao/713/2013(H7N9)
A/wild duck/Korea/SH19-47/2010(H7N9)
A/duck/Jiangxi/3096/2009(H7N9)
A/chicken/British Columbia/GSC human B/04(H7N3)
A/avian/Delaware Bay/226/2006(H7N3)
A/American green-winged teal/California/44242-906/2007(H7N3)
A/American black duck/New Brunswick/02490/2007(H7N3)
A/American black duck/NB/2538/2007(H7N3)
A/mallard/Postdam/178-4/1983(H2N2)
A/duck/Hong Kong/319/1978(H2N2)
A/emperor goose/Alaska/44297-260/2007(H2N2)
1.0
0.8
0.6
0.4
0.2
0.0

Fig. 7. Phylogenetic tree of segment 6 of Influenza A virus genomes based on Clustal Omega.

Steps:

1. Construct Chaos Game Representation (CGR) of each DNA sequence.
2. Compute discrete Fourier transform (DFT) of each corresponding CGR sequence.
3. Compute power spectrum of each DFT sequence, respectively.
4. Apply even scaling method to the set of Power Spectra, which all sequence is scaled to the maximum length *m* of the underlying data set.
5. Construct phylogenetic tree using UPGMA method

3. Results

Our method is compared to the state-of-the-art Clustal Omega for efficiency and accuracy. Phylogenies of our method and Clustal Omega method are drawn using Matlab 7.14 and MEGA 6 [28]. Computations in this research are implemented on a PC with configuration of Intel Core i7 CPU 2.40 GHz and 8 Gb RAM.

3.1. Human rhinovirus

HRV are associated with upper and lower respiratory diseases, and are the predominant cause of the common cold and cold-like illnesses. Palmenberg et al. [29] have classified the complete HRV genomes into three genetically distinct groups within the genus Enterovirus and the family Picornaviridae. The underlying dataset consists of three groups HRV-A, HRV-B, HRV-C of 113 genomes and three outgroup sequences HEV-C. While the genomes were classified correctly in Palmenberg's work, the running time was quite high due to the use of multiple sequence alignment to construct the evolutionary tree. We test our method on the same dataset in this paper. As illustrated in the phylogenetic tree in Fig. 4 the three groups of HRV are clearly separated and are distinguished from the outgroup HEV-C using our method. The running time was 7 s. We also clustered the dataset using Clustal Omega (Fig. 5). The genomes are classified into the correct groups, but it took Clustal Omega 19 min and 35 s to finish the classification.

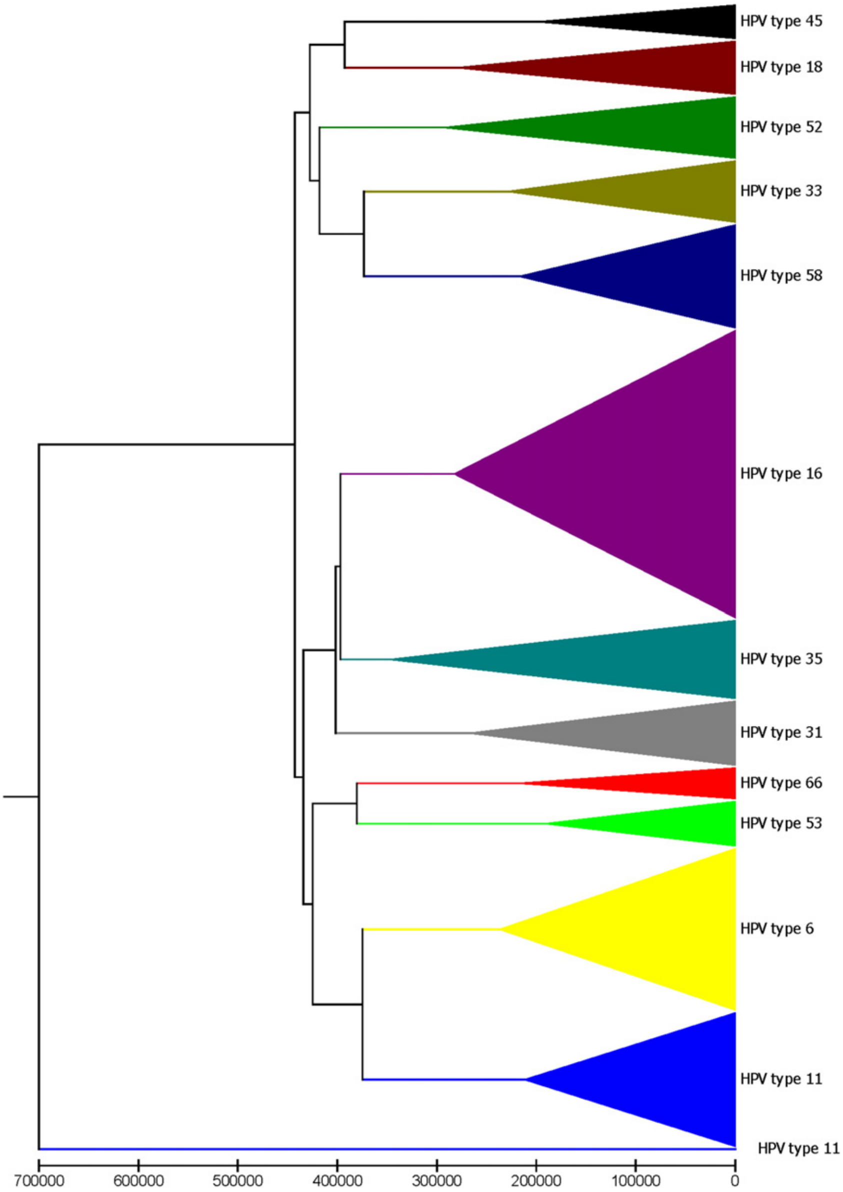

Fig. 8. Phylogenetic tree of 400 HPV genomes of 12 genotypes based on our method.

3.2. Influenza

Influenza A viruses have been a major health threat to both human society and animals [30]. Influenza A viruses are the most dangerous due to their wide natural host range, including birds, horses, swine, and humans; and they are known to have high degree of genetic and antigenic variability [31,32]. Some of the most lethal subtypes are H1N1, H2N2, H5N1, H7N3, and H7N9, which are responsible for many pandemic flus. Influenza A viruses are single-stranded, segmented RNA viruses, which are classified based on the viral surface proteins hemagglutinin (HA or H) and neuraminidase (NA or N) [33]. We examine segment 6 of Influenza A virus genome, which encodes for neuraminidase (NA), to test our method. As illustrated by the phylogenetic trees, the dataset is classified correctly into biological groups by our method in less than 1 s (Fig. 6), except for one case A/duck/Guangxi/030D/2009 (H1N1). It takes Clustal Omega 9 s for classification, however it misplaces the virus A/turkey/VA/505477-18/2007(H5N1) into H1N1 group, and a part of H1N1 viruses are incorrectly grouped with H5N1 subtype (Fig. 7).

3.3. HPV

Cervical cancer is the second most common cancer among women worldwide, mostly caused by Human Papillomavirus (HPV) [34]. Therefore, HPV is considered the most common sexually transmitting DNA virus. Low risk HPV types such as 6 and 11 can cause genital warts or benign. High risk HPV types such as 16 and 18 account for about 70% of cervical cancer [35]. Henceforth, it is important to not only classify HPV into high and low risk types, but also to identify HPV genotypes infecting the patients as quick as possible. As a result, multiple approaches were proposed to focus on predicting the HPV risk types, using various tools, such as text mining [36], support vector machines [37], decision tree [38], and ensembled support vector machines with protein secondary structures [39].

However, HPV genotypes are hardly detected for many cases such as inadequate samples or low amplification signals of some genotypes. Contamination with previously amplified material can also lead to false positive results [12]. One way to avoid these problems is to use some computational methods for identifying HPV types. By the special characteristics of CGR, partial DNA sequence can still display genome signatures of the whole sequence, make it possible to compare genomes when only partial genomes are available. Therefore, our proposed method would make a good candidate for classifying and predicting HPV genotypes quickly and efficiently.

In this work, we apply our method on the data set of 400 HPV genomes of 12 genotypes 6, 11, 16, 18, 31, 33, 35, 45, 52, 53, 58, and 66, used in [12]. Except for one genome, our method classifies the dataset into the correct biological groups in less than 30 s (Fig. 8). Even though Clustal Omega is able to classify the dataset correctly, its running time is 2 h and 17 min, which is significantly longer (Fig. 9).

4. Discussion and conclusion

Numerical representations of biological sequences have gained more and more attention in bioinformatics research. The advantage of this direction is that large sequence data can be handled via the use of digital signal processing method, e.g. with DFT, important features that are not visible in time domain can be easily detected in frequency domain. However, most of the existing numerical representation methods suffer from being not one-to-one, i.e. given the transformed numerical sequence, it might not be possible to reconstruct the original DNA sequence [4]. One main reason for this problem is that most of those representation methods are nucleotide mapping, i.e. each nucleotide is assigned a fixed numerical value. As a consequence, the natural history of the sequence cannot be preserved perfectly.

We propose to use Chaos Game Representation (CGR) in numerical representation of biological sequences. CGR is a one-to-one sequence mapping. That means the nature of the DNA sequence can be stored and maintained after the transformation. For any given point on the CGR numerical sequence, we can reconstruct the whole original DNA sequence up to that point [8]. None of the other numerical representation is able to do the same. Moreover, CGR can be employed as an alternative genomic signature, where it provides a unique visualization of a DNA sequence of any length in a two-dimensional plot. Various interesting fractal geometrical patterns, such as parallel lines, triangles, rectangles, and squares can be observed via the use of CGR, in which researchers can discover interesting properties of the DNA sequence that might not be detected by computers.

Several methods to classify genes and genomes are alignment-based, in which optimal alignments are obtained by using selected scoring systems. Even though these methods often give accurate classification of biological sequences, their main drawbacks are due to significantly high running time and memory consumption [3]. As a result, alignment-free technique has been extensively developed recently in order to provide more efficient methods [3,16,40].

In our research, by realizing 2D CGR coordinates as complex numbers, we combine the advantages of CGR and complex DFT to construct a new alignment-free method to classify genes and genomes. After

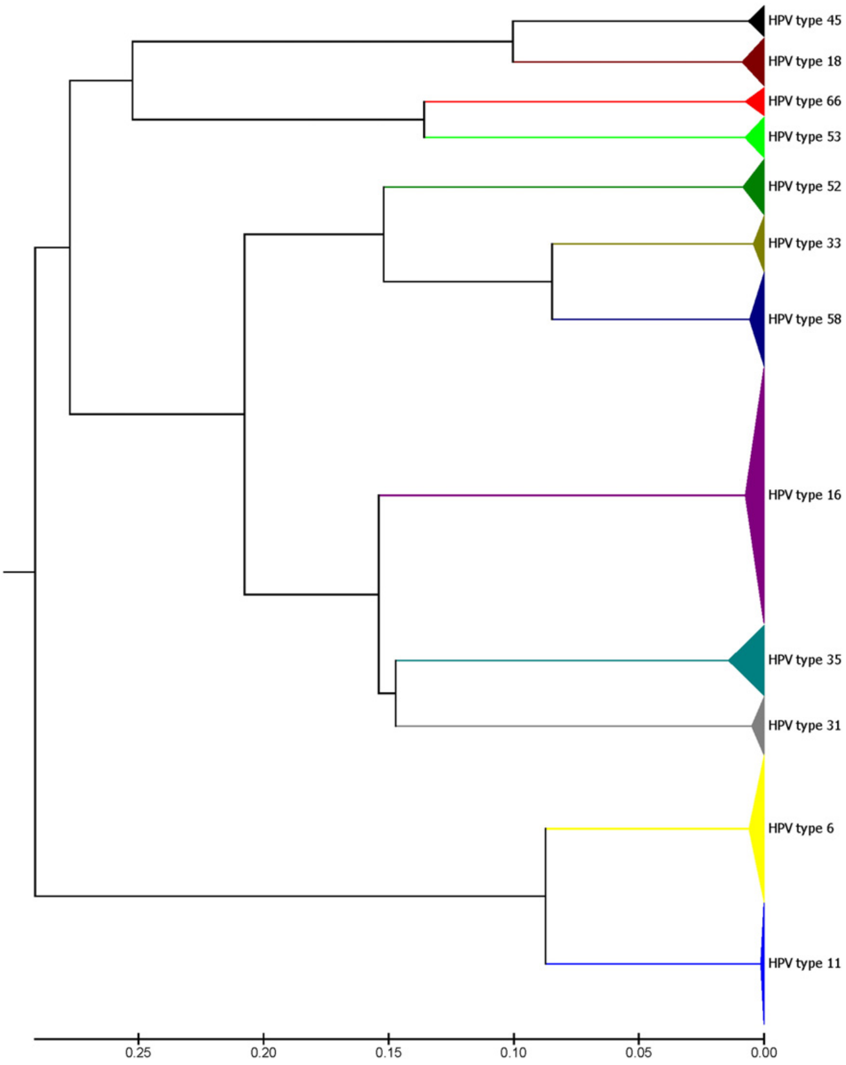

Fig. 9. Phylogenetic tree of 400 HPV genomes of 12 genotypes based on Clustal Omega.

constructing power spectra of DNA sequences [3], we propose to apply even scaling method [4] on power spectra so that sequences of different lengths can be embedded into the same Euclidean space.

Our method is compared to the state-of-the-art method Clustal Omega for efficiency and accuracy. While providing comparable biological classification, the running time for our method is significantly faster compared to Clustal Omega. Specifically, for large data set like HPV, it took our method less than 30 s to finish the classification, while the running time for Clustal Omega was 2 h and 17 min. Another concern when classifying and predicting HPV genotypes is that HPV genotypes are hardly detected in cases of inadequate samples or low amplification signals of some genotypes. This is the limitation of various kinds of HPV genotyping tests used in clinical laboratories [12]. Computational methods like our proposed method can resolve this limitation, as via the use of CGR, partial sequence can display genome signatures of the whole sequence, thus gives rise to possibility of comparing genomic sequences when only parts of the genomes are available [9]. The misplacement of one genome of HPV type 11 in our method will be justified in our future research.

Acknowledgment

This research is supported by the National Natural Sciences Foundation of China (31271408 to S.S.-T. Yau), Tsinghua University start-up fund (to S. S.-T. Yau) and Tsinghua University independent research project grant (to S. S.-T. Yau). The corresponding author would like to thank National Center for Theoretical Sciences (NCTS) for providing excellent research environment while part of this research was done.

Appendix A. Supplementary data

Supplementary material S1 contains the proof of Theorem 2.1, as well as tables of accession numbers and names of the underlying datasets (Table S1–S4). Supplementary data associated with this article can be found in the online version, at doi: http://dx.doi.org/10.1016/j.ygeno.2016.08.002.

References

[1] S. Bai Arniker, H.K. Kwan, Advanced numerical representation of DNA sequences, International Conference on Bioscience, Biochemistry and Bioinformatics IPCBEE 2012, p. 1.
[2] R.F. Voss, Evolution of long-range fractal correlations and 1/f noise in DNA base sequences, Phys. Rev. Lett. 68 (1992) 3805.
[3] T. Hoang, C. Yin, H. Zheng, C. Yu, R.L. He, S.S.-T. Yau, A new method to cluster DNA sequences using Fourier power spectrum, J. Theor. Biol. 372 (2015) 135–145.
[4] C. Yin, S.S.-T. Yau, An improved model for whole genome phylogenetic analysis by Fourier transform, J. Theor. Biol. 382 (2015) 99–110.
[5] C. Yin, S.S.-T. Yau, A Fourier characteristic of coding sequences: origins and a non-Fourier approximation, J. Comput. Biol. 12 (2005) 1153–1165.
[6] D. Anastassiou, Frequency-domain analysis of biomolecular sequences, Bioinformatics 16 (2000) 1073–1081.
[7] S.S.-T. Yau, J. Wang, A. Niknejad, C. Lu, N. Jin, Y.-K. Ho, DNA sequence representation without degeneracy, Nucleic Acids Res. 31 (2003) 3078–3080.
[8] H.J. Jeffrey, Chaos game representation of gene structure, Nucleic Acids Res. 18 (1990) 2163–2170.
[9] P.J. Deschavanne, A. Giron, J. Vilain, G. Fagot, B. Fertil, Genomic signature: characterization and classification of species assessed by chaos game representation of sequences, Mol. Biol. Evol. 16 (1999) 1391–1399.
[10] J.S. Almeida, J.A. Carrico, A. Maretzek, P.A. Noble, M. Fletcher, Analysis of genomic sequences by chaos game representation, Bioinformatics 17 (2001) 429–437.
[11] J. Joseph, R. Sasikumar, Chaos game representation for comparison of whole genomes, BMC Bioinf. 7 (2006) 243.
[12] W. Tanchotsrinon, C. Lursinsap, Y. Poovorawan, A high performance prediction of HPV genotypes by chaos game representation and singular value decomposition, BMC Bioinf. 16 (2015) 71.
[13] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (2004) 1792–1797.
[14] K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Res. 30 (2002) 3059–3066.
[15] M.A. Larkin, G. Blackshields, N. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, et al., Clustal W and Clustal X version 2.0. Bioinformatics 23 (2007) 2947–2948.
[16] S. Vinga, J. Almeida, Alignment-free sequence comparison - a review, Bioinformatics 19 (2003) 513–523.
[17] S.S.-T. Yau, C. Yu, R. He, A protein map and its application, DNA Cell Biol. 27 (2008) 241–250.
[18] C. Yu, M. Deng, S.S.-T. Yau, DNA sequence comparison by a novel probabilistic method, Inf. Sci. 181 (2011) 1484–1492.
[19] A. Pandit, S. Sinha, Using genomic signatures for HIV-1 sub-typing, BMC Bioinf. 11 (2010) S26.
[20] B.E. Blaisdell, A measure of the similarity of sets of sequences not requiring sequence alignment, Proc. Natl. Acad. Sci. 83 (1986) 5155–5159.
[21] J. Wen, R.H. Chan, S.-C. Yau, R.L. He, S.S. Yau, K-mer natural vector and its application to the phylogenetic analysis of genetic sequences, Gene 546 (2014) 25–34.
[22] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, R. Ramaswamy, Prediction of probable genes by Fourier analysis of genomic sequences, Bioinformatics 13 (1997) 263–270.
[23] D. Kotlar, Y. Lavner, Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions, Genome Res. 13 (2003) 1930–1937.
[24] C. Yin, S.S.-T. Yau, Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence, J. Theor. Biol. 247 (2007) 687–694.
[25] R.R. Sokal, A statistical method for evaluating systematic relationships, Univ. Kans. Sci. Bull. 38 (1958) 1409–1438.
[26] F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, et al., Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Mol. Syst. Biol. 7 (2011) 539.
[27] A.V. Oppenheim, R.W. Schafer, J.R. Buck, et al., Discrete-Time Signal Processing, Prentice-hall Englewood Cliffs, 1989.
[28] K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30 (2013) 2725–2729.
[29] A.C. Palmenberg, D. Spiro, R. Kuzmickas, S. Wang, A. Djikeng, J.A. Rathe, C.M. Fraser-Liggett, S.B. Liggett, Sequencing and analyses of all known human rhinovirus genomes reveal structure and evolution, Science 324 (2009) 55–59.
[30] D.J. Alexander, A review of avian influenza in different bird species, Vet. Microbiol. 74 (2000) 3–13.
[31] R.J. Garten, C.T. Davis, C.A. Russell, B. Shu, S. Lindstrom, A. Balish, W.M. Sessions, X. Xu, E. Skepner, V. Deyde, et al., Antigenic and genetic characteristics of swine-origin 2009 A (H1N1) influenza viruses circulating in humans, Science 325 (2009) 197–201.
[32] P. Palese, J.F. Young, Variation of influenza A, B, and C viruses, Science 215 (1982) 1468–1474.
[33] R.G. Webster, W.J. Bean, O.T. Gorman, T.M. Chambers, Y. Kawaoka, Evolution and ecology of influenza A viruses, Microbiol. Rev. 56 (1992) 152–179.
[34] M. Arbyn, X. Castellsague, S. De Sanjose, L. Bruni, M. Saraiya, F. Bray, J. Ferlay, Worldwide burden of cervical cancer in 2008, Ann. Oncol. 22 (2011) 2675–2686.
[35] J.S. Smith, L. Lindsay, B. Hoots, J. Keys, S. Franceschi, R. Winer, G.M. Clifford, Human papillomavirus type distribution in invasive cervical cancer and high-grade cervical lesions: a meta-analysis update, Int. J. Cancer 121 (2007) 621–632.
[36] S.-B. Park, S. Hwang, B.-T. Zhang, Classification of human papillomavirus (HPV) risk type via text mining, Genomics Inform. 1 (2003) 80–86.
[37] S. Kim, B.-T. Zhang, Human papillomavirus risk type classification from protein sequences using support vector machines, in: F. Rothlauf, J. Branke, S. Cagnoni, E. Costa, C. Cotta, R. Drechsler, E. Lutton, P. Machado, J.H. Moore, J. Romero, G.D. Smith, G. Squillero, H. Takagi (Eds.), Applications of Evolutionary Computing, Lecture Notes in Computer Science, Springer, Berlin Heidelberg 2006, pp. 57–66.
[38] S.-B. Park, S. Hwang, B.-T. Zhang, Classification of the risk types of human papillomavirus by decision trees, in: J. Liu, Y. Cheung, H. Yin (Eds.), Intelligent Data Engineering and Automated Learning, Lecture Notes in Computer Science, Springer, Berlin Heidelberg 2003, pp. 540–544.
[39] S. Kim, J. Kim, B.-T. Zhang, Ensembled support vector machines for human papillomavirus risk type prediction from protein secondary structures, Comput. Biol. Med. 39 (2009) 187–193, http://dx.doi.org/10.1016/j.compbiomed.2008.12.005.
[40] C. Yu, T. Hernandez, H. Zheng, S.C. Yau, H.H. Huang, R.L. He, J. Yang, S.S.T. Yau, Real time classification of viruses in 12 dimensions, PloS one 8 (5) (2013) p.e64328.