Molecular Phylogenetics and Evolution 141 (2019) 106633

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Fast and accurate genome comparison using genome images: The Extended Natural Vector Method

Shaojun Pei[a,1], Wenhui Dong[a,1], Xiuqiong Chen[a], Rong Lucy He[b], Stephen S.-T. Yau[a,*]

[a] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, PR China
[b] Department of Biological Sciences, Chicago State University, Chicago, IL 60628, USA

ARTICLE INFO

Keywords:
Chaos game representation
Extended natural vector
Genome comparison

ABSTRACT

Using numerical methods for genome comparison has always been of importance in bioinformatics. The Chaos Game Representation (CGR) is an effective genome sequence mapping technology, which converts genome sequences to CGR images. To each CGR image, we associate a vector called an Extended Natural Vector (ENV). The ENV is based on the distribution of intensity values. This mapping produces a one-to-one correspondence between CGR images and their ENVs. We define the distance between two DNA sequences as the distance between their associated ENVs. We cluster and classify several datasets including *Influenza* A viruses, *Bacillus* genomes, and *Conoidea* mitochondrial genomes to build their phylogenetic trees. Results show that our ENV combining CGR method (CGR-ENV) compares favorably in classification accuracy and efficiency against the multiple sequence alignment (MSA) method and other alignment-free methods. The research provides significant insights into the study of phylogeny, evolution, and efficient DNA comparison algorithms for large genomes.

1. Introduction

Genome sequence analysis is considered an indispensable part of the field of bioinformatics for understanding the evolution of species and for describing and understanding the evolution of species. This area has been developing rapidly in recent years (Ackermann and Kropinski, 2007; Moore et al., 2011; Vinga and Almeida, 2003), and has been applied to molecular phylogeny, comparative genomics, gene prediction and annotation. Many methods have been proposed to compare genome sequences.

Most traditional methods are alignment-based methods. The comparison of more than two sequences is known as Multiple Sequence Alignment (MSA). The most common multiple sequence alignment methods are ClustalW (Higgins et al., 1994), MUSCLE (Edgar, 2004), T-Coffee (Notredame et al., 2000) and MAFFT (Katoh and Kuma, 2002). MSA methods involve computing similarity scores between DNA or amino acid sequences. They have often been used to predict phylogenetic trees using a single gene or multiple conserved genes. Unfortunately, it is not uncommon for the phylogenetic tree obtained from focusing on one gene or set of genes to differ from the tree obtained from focusing on another gene or set of genes (Ludwig et al., 1998; Lang et al., 2013). Because of this, selecting a set of genes appropriate for the number and diversity of the taxa is very important (Wang and Wu, 2013). To improve the accuracy of phylogenetic trees, performing MSA with longer sets of genome sequences is necessary (Brown et al., 2001). But MSA methods are very time-consuming and quite expensive in memory usage. With a sharp increase in the number of biological genome sequences, these traditional sequence alignment methods become unworkable. To be able to compare entire genome sequences, alignment-free methods need to be used. In contrast to the traditional MSA methods, alignment-free methods tend to be computationally efficient and can utilize all of the genomic information (Vinga and Almeida, 2003). For example, in (Deng et al., 2011), the alignment-free Natural Vector (NV) method is proposed. A natural vector describing the distribution (i.e., the locations and number of occurrences) of each nucleotide, is associated to each DNA sequence. The NV method is unsupervised and does not require manual intervention. It results in a one-to-one correspondence between a DNA sequence and its associated natural vector, which reflects the biological properties of the original genome sequences.

The Chaos Game Representation (CGR) is an iterative mapping technique that divides genome sequences into certain units, and then finds the correlation between their positions in the gene sequence (Jeffrey, 1990). Specifically, this technique assigns each oligonucleotide in a DNA sequence to a position in the plane. This mapping allows the DNA sequence to be depicted as a CGR image. Image processing

* Corresponding author.
E-mail address: yau@uic.edu (S.S.-T. Yau).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ympev.2019.106633
Received 4 April 2019; Received in revised form 25 September 2019; Accepted 25 September 2019
Available online 26 September 2019
1055-7903/

related methods can be used to find the patterns in the genome sequences (Almeida et al., 2001; Deschavanne et al., 1999). For example, in (Ni et al., 2018), the mean structural similarity (MSSIM) coefficient between pairs of CGR images is used to measure the degree of similarity between the corresponding genomes. However, most of these methods extract certain features from the CGR images, but these features do not contain all the information in the images, which means it is not possible to fully recover a CGR image from the set of features that are extracted.

In this paper, we propose a novel method to compare the two-dimensional CGR image matrices. To each CGR image, we associate a vector, called Extended Natural Vector (ENV), that describes the numbers and distributions of intensity values in the CGR images. We prove that the ENV defined in this context can distinguish image matrices in a strict one-to-one fashion. A natural distance between two genome sequences is the distance between the corresponding extended natural vectors of their CGR images. Combining CGR with ENV (CGR-ENV) method, we propose a new numerical vector approach to find the similarities and differences between genome sequences. Tested on datasets of *Influenza* A viruses, *Bacillus* genomes and *Conoidea* mitochondrial genomes, we find that this novel approach gives better results than ClustalW, NV, MSSIM-combined CGR methods and Higher-Order Markov model (Yang et al., 2016).

2. Materials and methods

2.1. Materials

We apply our method for phylogeny reconstruction three datasets, each consisting of whole genome sequences. The GenBank IDs are shown in Tables S1-S3.

Dataset 1 contains 27 *Influenza* A viruses. According to the reference taxonomy in NCBI, these 27 *Influenza* A viruses are composed of 5 subtypes (including 8 H5N1, 6 H1N1, 2 H2N2, 5 H7N3 and 6 H7N9 respectively).

Dataset 2 contains 36 *Bacillus* genomes. According to the taxonomy based on the classification of the NCBI, the population of 36 *Bacillus* consists of 2 orders (*Bacillales* and *Lactobacillales*), 7 families (*Alicyclobacillaceae, Bacillaceae, Staphylococcaceae, Listeriaceae, Leuconostocaceae, Lactobacillaceae, Streptococcaceae*).

Dataset 3 contains 9 *Conoidea* mitochondrial genomes. The list of *Conoidea* superfamily analyzed in this study is corresponding to families *Conidae, Borsoniidae, Mangeliidae, Clathurellidae, Clavatulidae, Turridae, Terebridae*.

2.2. Methods

In this study, a novel method which we call the CGR-ENV method is applied to genome sequence data. In this method, each genome sequence is first converted into a CGR image using the method proposed by Jeffrey (Jeffrey, 1990).

2.2.1. Chaos Game Represntation (CGR) images

The CGR method transforms DNA sequences to images in order to reveal patterns in the sequences. CGR is defined iteratively by Eq. (1). For a DNA sequence $s_1 s_2, ..., s_n ...$, the corresponding CGR position $X_n = (x_n, y_n)$ is given by:

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

where W is (0,0) if s_n is A, (0,1) if s_n is C, (1,1) if s_n is G, or (1,0) if s_n is T.

Basically, each pixel in the CGR images is associated with the frequency of a specific word (Fig. 1). Overlaying a CGR image with a grid of appropriate size, the frequency of each word is equal to the number of occurrences in its associated grid position. In order to obtain the frequency matrix of n-string words, a $2^n \times 2^n$ grid must be used

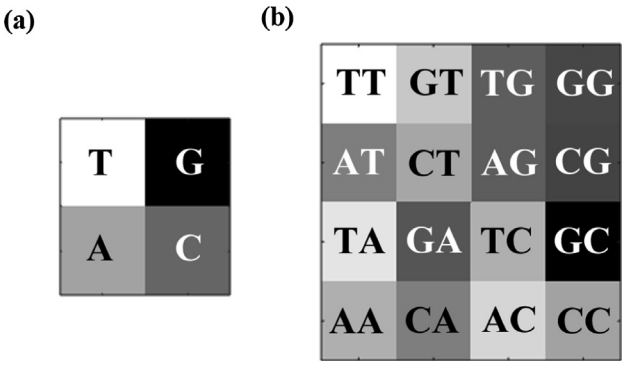


Fig. 1. (a) The 2×2 CGR image of A/mallard/Nova Scotia/00088/2010(H1N1) with words length $n = 1$. The frequencies of one letter are represented by the gray scale. (b) The 4×4 CGR image of A/mallard/Nova Scotia/00088/2010(H1N1) with words length $n = 2$. The frequencies of two letters are represented by the gray scale.

(Almeida et al., 2001). We let the gray value of the greatest frequency grid be 255, and the lowest frequency grid be 0. Then we can obtain a grayscale CGR image. Word frequencies are displayed by the intensity of each pixel.

For example, Fig. 2a is the CGR image of 'A/mallard/Nova Scotia/00088/2010(H1N1)' with words length $n = 3$ and Fig. 2b is the distribution of its gray values. Qualitative and quantitative expressions of the order, regularity, structure, and complexity of DNA sequences are obtained from the CGR image, which simultaneously displays both local and global patterns of the sequence (Deschavanne et al., 1999).

After the CGR images are produced, the ENVs corresponding to the CGR images are constructed.

2.2.2. Construction of the ENV on a two-dimensional image pixel matrix

A grayscale CGR image having $2^n \times 2^n$ pixels, with 256 possible intensities at each pixel (i.e., with 8 bits there are $2^8 = 256$ possible values) determines a $2^n \times 2^n$ matrix Q with entries $q(i, j) \in K = \{0, 1, 2, 3, ..., 255\}$, where (i, j) represents the locations of gray values in matrix.

Denote the cardinality of $q^{-1}(k)$ by n_k, where q^{-1} is the inverse mapping of q. Thus for $k \in K = \{0, 1, 2, ..., 255\}$, n_k is the total number of k-valued pixels in the gray-scale image. Then $N = 2^n \times 2^n = \sum_{k=0}^{255} n_k$ is the total number of pixels. Furthermore, let $q^{-1}(k) = \{(i_{s,k}, j_{s,k}) | s = 1, 2, ..., n_k\}$ be the set of all k-valued pixel positions. In the example matrix (Fig. 2b), $q^{-1}(215)$ consists of four pixels. The value $k = 215$ shows up as following

$$\{(i_{1,215}, j_{1,215}), (i_{2,215}, j_{2,215}), (i_{3,215}, j_{3,215}), (i_{4,215}, j_{4,215})\}$$
$$= \{(1, 6), (2, 5), (2, 8), (6, 7)\}.$$

Thereafter, the ENV can be determined by forming a vector from the following group of quantities.

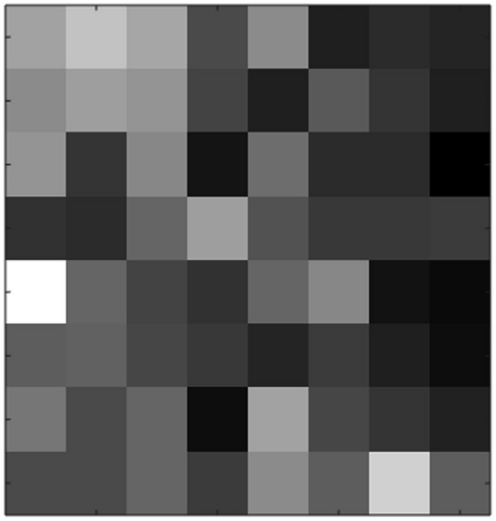
1. The first group of components in the ENV are the 256 counts $(n_0, n_1, ..., n_{255})$. They're all non-negative integers bounded by the size of the matrix. Some of them may be zero. In our example above, $n_{215} = 4$.
2. The second group of components in the ENV are mean pixel locations $\vec{\mu}_k = (\mu_{1,k}, \mu_{2,k})$ for intensity values $k = 0, 1, ..., 255$.

$$\mu_{1,k} := \frac{\sum_{s=1}^{n_k} i_{s,k}}{n_k} = \frac{i_{1,k} + i_{2,k} + \cdots + i_{n_k,k}}{n_k}, \tag{2}$$

$$\mu_{2,k} := \frac{\sum_{s=1}^{n_k} j_{s,k}}{n_k} = \frac{j_{1,k} + j_{2,k} + \cdots + j_{n_k,k}}{n_k}, \tag{3}$$

where $(i_{s,k}, j_{s,k}) \in q^{-1}(k)$, for all $s = 1, 2, ..., n_k$.

(a)



(b)

78	49	73	161	98	215	196	205
98	83	88	171	215	147	186	215
88	186	102	230	127	196	196	255
191	196	132	83	152	181	181	176
0	132	171	191	132	102	235	245
142	137	166	181	205	176	215	240
117	161	132	240	78	166	186	210
161	161	132	176	98	142	39	142

Fig. 2. (a) The 8×8 CGR image of A/mallard/Nova Scotia/00088/2010(H1N1) with words length $k = 3$. (b) The distribution of gray values is listed in the table.

Let us look at our example, where $\vec{\mu}_{215} = (\mu_{1,215}, \mu_{2,215}) = (\frac{1+2+2+6}{4}, \frac{6+5+8+7}{4}) = (\frac{11}{4}, \frac{26}{4})$. Define $\vec{\mu}_k := (0, 0)$ if $n_k = 0$. For all other k's, these vector components are positive rational numbers and not necessarily integers. The ENV has components from $\vec{\mu}_0$ through $\vec{\mu}_{255}$.

3. The third group of parameters that we include in the ENV are the normalized higher order central moments. There is a set of D's for each $k \in K$. Let $D_{k,0,0} := 0$ for $k \in K$. For any other exponent pair (r, s), we define

$$D_{k,r,s} := \sum_{t=1}^{n_k} \frac{(i_{t,k} - \mu_{1,k})^r \cdot (j_{t,k} - \mu_{2,k})^s}{(n_k)^{r+s} \cdot N^{r+s-1}} \tag{4}$$

where $k \in K$, r is an arbitrary non-negative integer, and $s = 0, 1, 2, \ldots, n_k$.

Plugging $r = 0$ and $s = 0$ into the Eq. (4) yields only normalized counts $n_k \cdot N$ which are already captured. $D_{k,1,0}$ and $D_{k,0,1}$ turn out to be zero, a property of the location means $\mu_{1,k}$ and $\mu_{2,k}$. Thus central moments of combined degree 0 and 1 can be omitted.

cwith D's ordered lexicographically starting at $k = 0$ with degree two.

$$\begin{aligned} &< n_0, \mu_{1,0}, \mu_{2,0}, D_{0,0,2}, D_{0,1,1}, D_{0,2,0}, \ldots \\ &\qquad n_1, \mu_{1,1}, \mu_{2,1}, D_{1,0,2}, D_{1,1,1}, D_{1,2,0}, \ldots \\ &\qquad \ldots \\ &n_{255}, \mu_{1,255}, \mu_{2,255}, D_{255,0,2}, D_{255,1,1}, D_{255,2,0}, \ldots > \end{aligned} \tag{5}$$

Theorem 1. Suppose the entries of a two-dimensional $m \times n$ matrix are elements of the finite set $K = \{0, 1, 2, \ldots, 255\}$ then the corresponding ENV determines all the matrix entries.

From Theorem 1, we can see that the information in the ENV is enough to theoretically determine the entire grayscale image matrix. The proof of Theorem 1 is given in Appendix A.

Obviously, higher central moments converge to zero for a random generated distribution matrix since for any given k,

$$D_{k,r,s} = \sum_{t=1}^{n_k} \frac{(i_{t,k} - \mu_{1,k})^r \cdot (j_{t,k} - \mu_{2,k})^s}{(n_k)^{r+s} \cdot N^{r+s-1}} \leqslant \sum_{t=1}^{n_k} \frac{\max\limits_{\{t \in 1,2,\ldots,n_k\}} |i_{t,k}|^r \cdot |j_{t,k}|^s}{(n_k)^{r+s} \cdot N^{r+s-1}} \leqslant \frac{\max\limits_{\{t \in 1,2,\ldots,n_k\}} |i_{t,k}|^r \cdot |j_{t,k}|^s}{(n_k)^{r+s-1} \cdot N^{r+s-1}} \leqslant \frac{N^r \cdot N^s}{(n_k)^{r+s-1} \cdot N^{r+s-1}} = \frac{N}{(n_k)^{r+s-1}}, \tag{6}$$

where $N = \sum_{k=0}^{255} n_k$.

It is clear that $n_k \geqslant 1$, otherwise, there is no any k-th grayscale distributed in the matrix.

From the viewpoint of probability, suppose that the expected value of any grayscale value is $n_k = N/256$ (uniform distribution) for an image with N values from the given distribution matrix.

In our case, for the given distribution matrix in $2^n \times 2^n$, we naturally get the total entries of the matrix as $N := 2^n \times 2^n$. Therefore,

$$\lim_{r+s} \frac{N}{(n_k)^{r+s-1}} = \lim_{r+s} \frac{N}{(N/256)^{r+s-1}} = \lim_{r+s} \frac{256^{r+s-1}}{N^{r+s-2}}. \tag{7}$$

Clearly, this limit tends to 0 as $r + s$ approaches n_k. Specifically, in our example, $D_{215,1,2} = -0.00073$, i.e., higher normalized central moments starting from 3^{rd} moment will converge to 0. So for each CGR image, we used the 1536-dimensional ENV listed in (8) in our experiments.

$$\begin{aligned} &< n_0, \ n_1, \ n_2, \ \ldots, n_{255}, \\ &\qquad \mu_0^1, \mu_1^1, \ \mu_{A_2}^1, \ \ldots, \mu_{255}^1, \\ &\qquad \mu_0^2, \mu_1^2, \mu_{A_2}^2, \ldots, \mu_{255}^2, \\ &D_{0,2,0}, D_{1,2,0}, D_{2,2,0}, \ldots, D_{255,2,0}, \\ &D_{0,0,2}, D_{1,0,2}, D_{2,0,2}, \ldots, D_{255,0,2}, \\ &D_{0,1,1}, D_{1,1,1}, D_{2,1,1}, \ldots, D_{255,1,1} > . \end{aligned} \tag{8}$$

Algorithm 1 below shows the whole procedure.

Algorithm 1. CGR-ENV method

Input: a DNA sequence
Output: the Extended Natural Vector
1: Given the fixed word length n
2: Generate the grayscale Chaos Game Representation (CGR) image of the DNA sequence
3: Calculate the Extended Natural Vector (ENV) of grayscle CGR image based on the quantities of (8)

2.2.3. Distance measure

For two different genome sequences S_1 and S_2, we can describe them with the ENV $V_1 = (p_1, p_2, \ldots, p_Q)$ and $V_2 = (q_1, q_2, \ldots, q_Q)$ in Q-dimensional space, where $Q = 256 \times 6 = 1536$, p_i and q_i are listed in (8). The Euclidean distance between V_1 and V_2 is used as the evolutionary distance between two corresponding genome sequences in the present approach.

To evaluate our method, we choose three alignment-free methods

including the NV method, the MSSIM combined-CGR method and the One-dimensional CGR-combined Higher-Order Markov Model as comparisons.

2.3. Other methods

2.3.1. Natural Vector (NV) method

Let $S = s_1 s_2 ... s_n$ be a DNA sequence of length n and $s_i \in A, C, G, T, i = 1, 2, ..., n$. For $K \in \{A, C, G, T\}$, we define $w_K(\cdot)$: $\{A, C, G, T\} \rightarrow \{0, 1\}$ such that $w_K(s_i) = 1$ if $s_i = K$, otherwise $w_K(s_i) = 0$.

1. Let $n_K = \sum_{i=1}^{n} w_K(s_i)$ denote the number of nucleotide K in the DNA sequence S.
2. Let $\mu_K = \sum_{i=1}^{n} \frac{i \cdot w_K(s_i)}{n_K}$ be the mean position of nucleotide K.
3. Let $D_2^K = \sum_{i=1}^{n} \frac{(i-\mu_K)^2 w_K(s_i)}{n n_K}$ be a scaled variance of positions of nucleotide K.

The 12-dimensional NV of a DNA sequence S is defined by $(n_A, n_C, n_G, n_T, \mu_A, \mu_C, \mu_G, \mu_T, D_2^A, D_2^C, D_2^G, D_2^T)$.

2.3.2. Mean structural similarity (MSSIM) combined-CGR method

For a gray-scale image x, we define the average gray level α_x as:

$$\alpha_x = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{9}$$

where x_i is the gray value of image x, N is the size of image.

The standard deviation β_x is defined as

$$\beta_x = \left(\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \alpha_x)^2 \right)^{\frac{1}{2}} \tag{10}$$

And for two gray-scale images x, y, the MSSIM index is as follows:

$$MSSIM(x, y) = \frac{4\alpha_x \alpha_y \beta_x \beta_y}{(\alpha_x^2 + \alpha_y^2)(\beta_x^2 + \beta_y^2)} \tag{11}$$

2.3.3. One-dimensional CGR-combined Higher-Order Markov Model

The One-dimensional CGR can map a DNA sequence S of length m to a numeric sequence N according to the following equation:

$$N_i = (N_{i-1} + P_i)/4, \; N_0 = 0.5 \tag{12}$$

where P_i is 0, 1, 2, 3 for A, C, G and T respectively. Each n-string word is mapped into a sub-interval of width 4^{-n}, which means the number of N_i in each interval is equal to the number of each k-string word in the DNA sequence.

Then the higher order Markov model is used to characterize the DNA sequence S. A Markov model of order n represented in the form of a 4^{n+1} vector is denoted by M_n. Each element is the conditional possibility of $b_1 b_2 ... b_n b$:

$$P(b|b_1 b_2 ... b_n) = \frac{n(b_1 b_2 ... b_n b)}{\sum_{a \in \{A,C,G,T\}} n(b_1 b_2 ... b_n a)} \tag{13}$$

where $n(b_1 b_2 ... b_n a)$ is the number of the words $b_1 b_2 ... b_n a$ in the DNA sequence.

The distance between two vectors $V_1 = (p_1, ..., p_{4^{n+1}})$ and $V_2 = (q_1, ..., q_{4^{n+1}})$ is

$$D(V_1, V_2) = \frac{1}{2}\left(1 - \frac{\sum p_i q_i}{(\sum p_i^2)^{\frac{1}{2}} (\sum q_i^2)^{\frac{1}{2}}}\right) \tag{14}$$

2.4. The model of phylogenetic tree

The distance matrices of our CGR-ENV method, the NV method and the MSSIM method are calculated by using Matlab 2016a. The source code of the One-dimensional CGR-combined Higher-Order Markov model can be downloaded from (Yang et al., 2016). Based on the distance matrices, we use FastME which provides distance algorithms to infer phylogenies (Lefort et al., 2015). According to balanced minimum evolution, which is the very principle of Neighbor-Joining (NJ), FastME first constructs an initial tree based on NJ. Then topological moves, such as Nearest Neighbor Interchanges (NNIs) and Subtree Pruning and Regrafting (SPR) are performed to improve the structure of the tree. At last, we can get the Minimum Evolution (ME) phylogenetic tree.

For the MSA methods, we choose the ClustalW method to align the DNA sequences on the default parameters first (Higgins and Sharp, 1988). Then the Maximum Likelihood (ML) phylogenetic trees (Felsenstein, 1981) are constructed by Mega 7 (Kumar et al., 2016) using the GTR + G + I model of evolution and performing 100 bootstrap replicates (BP).

In addition, we apply the online tool EvolView to visualize and annotate phylogenetic trees (He et al., 2016). All the computation environment is an Intel(R) Core (TM) i7-7560U CPU @2.40 GHz Windows10 PC with 16.00 GB RAM.

3. Results

3.1. Influenza A viruses

Influenza A viruses are negative-sense, single-stranded, segmented RNA viruses which are a constant threat to both human and animal health because of their high mutation rate (Pei et al., 2019). They are classified based on the type of two surface glycoproteins: hemagglutinin (HA) and neuraminidase (NA) (Webster et al., 1992). The dataset used in this work consists of 27 viruses of the most lethal subtypes of *Influenza* A viruses, such as H1N1, H2N2, H5N1, H7N9, and H7N3.

According to the principle proposed by Sims (Sims et al., 2009), for n-string ($n > 1$), the n with maximum information is empirically determined but may be closely approximated by $n = \log_4 L$, where L is the average length of sequences. So we considered the words' length $n = 7$, and then obtained a 128 × 128 pixel CGR grayscale image. The ME phylogenetic trees of four methods (CGR-ENV method, NV method, MSSIM method, ClustalW method and High-Order Markov model) are shown in Fig. 3 and Figure S1-S4 respectively. From Fig. 3, we can see that our method can classify these *Influenza* viruses correctly into five groups, which are consistent with the biological taxonomy. In contrast, the NV method divides H7N3 into two different branches (Figure S1).

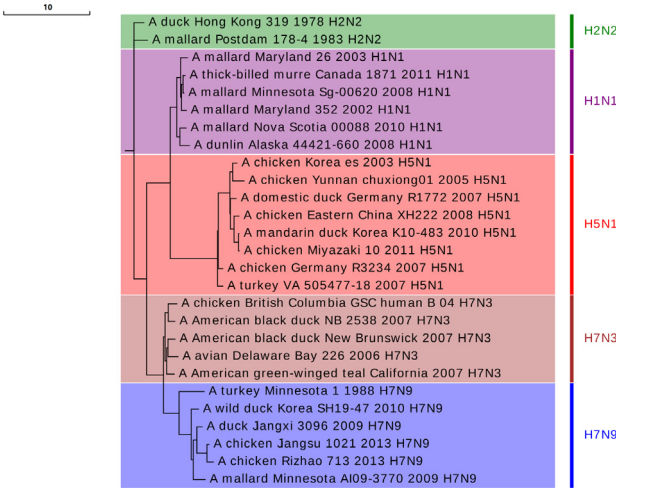


Fig. 3. Phylogenetic tree of 27 influenza A viruses 5 clusters: H1N1 (purple), H2N2 (green), H7N9 (blue), H7N3 (brown), H5N1 (red) by CGR-ENV method. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

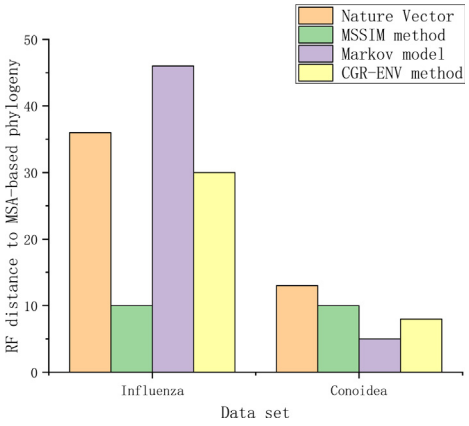


Fig. 4. The Robinson-Foulds distance of *Influenza* viruses and *Conoidea* species.

As shown in Fig. 4, the Robinson-Foulds (RF) distance between the phylogenetic trees generated by our method and the MSA method is smaller than the distance between those generated by the NV method. Although the phylogenetic tree constructed by the MSSIM method is closer to that by the ClustalW method, the 'A/turkey/VA/505477-18/2007 (H5N1)' misplaces in the branch of H1N1 in the phylogenetic tree constructed as shown in Figure S2 and S3.

3.2. Bacillus

Bacillus is a taxonomic class of bacteria that includes two orders, *Bacillales* and *Lactobacillales*. It contains several well-known pathogens such as *Bacillus anthracis*, which is the cause of anthrax. The data set in this study consists of seven different families, i.e. *Alicyclobacillaceae, Bacillaceae, Staphylococcaceae, Listeriaceae, Leuconostocaceae, Lactobacillaceae* and *Streptococcaceae*, 36 complete genome sequences in total. Fig. 5a shows the ME phylogenetic tree of these 36 *Bacillus* genome sequences computed by the CGR-ENV method with $n = 8$. We can see that these 36 *Bacillus* whole genome sequences are correctly classified according to the different Genotypes, where each family is marked with different colors. Meanwhile, the NV method, the MSSIM method and the Higher-Order Markov model don't perform well on this data set, as shown in Fig. 5b, Fig. 5c and Figure S5, since the sequences of *Streptococcaceae* are shuffled in the phylogenetic tree. The sequence length of our data set is more than 2 Mb, so based on the complexity of the ClustalW method, it may take more than 100 h (Higgins et al., 1994).

3.3. Conoidea mitochondrial genomes

Conoidea is a superfamily of marine mollusks and predatory sea snails which contains about 340 genera and subgenera. One authority considers that it includes approximately 4000 named living species (Puillandre et al., 2008), such as the terebras, the turrids (also named as auger shells or auger snails) and the cones. But the phylogeny of this superfamily is poorly analyzed and several families are thought to be polyphyletic. This dataset is used to aim to confirm the main deep lineages reported within *Conidea* (Uribe et al., 2017) including three genera within *Conidae*, three closely related families (*Borsoniidae*, *Mangeliidae*, *Clathurellidae*) and three outgroups (*Terebridae*, *Turridae* and *Clavatulidae*). As shown in Fig. 6, the result obtained by our method is consistent with previous molecular phylogenies (Uribe et al., 2017), in which the genera *Profundiconus* and *Lilliconus* is recovered as a sister group to the remaining members of *Conidae*. Although the phylogenetic tree obtained using the NV method is similar, three outgroups are not divided into the outside branch (Fig. 6b). And the sequences of *Conidae* family don't cluster together in one branch (Fig. 6c-6d and Figure S6) by the MSSIM method, the ClustalW method and the Higher-Order Markov model.

3.4. Time statistics and algorithm complexity

Table 1 displays the times that the CGR-ENV method, the NV method, the MSSIM method, the ClustalW method and the Higher-Order Markov model to process the three datasets. We do all the calculations on the same machine and empty the memory to avoid redundancy and influence before each calculation.

From Table 1, we can conclude that the alignment-free methods are much more time-efficient than the ClustalW method. If we denote L as the length of genome sequence and k as the number of sequences, the complexity of our proposed CGR-ENV method is $O(kL^2)$, while the complexity of the ClustalW method is $O(k^2L^2)$. Because our method consists of two steps: one is to convert the sequences to grayscale images by CGR and the other is to identify the images by ENV, our

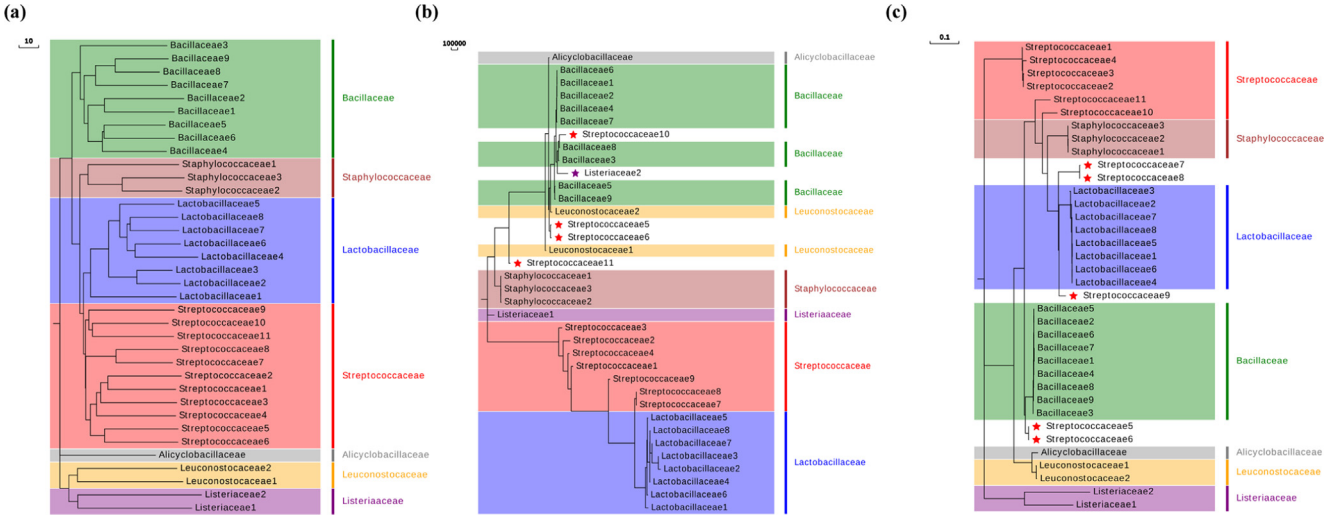


Fig. 5. The phylogenetic trees of 36 *Bacillus* genomes: *Alicyclobacillaceae* (gray), *Bacillaceae* (green), *Staphylococcaceae* (brown), *Listeriaceae* (purple), *Leuconostocaceae* (yellow), *Lactobacillaceae* (blue) and *Streptococcaceae* (red). Red and purple stars represent bacteria which aren't clustered with *Streptococcaceae* and *Listeriaceae* respectively. (a) The CGR-ENV method. (b) The NV method. (c) The MSSIM-combined CGR method. The sequences of *Streptococcaceae* are not classified in one branch of the phylogenetic tree by the NV method and the MSSIM method. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

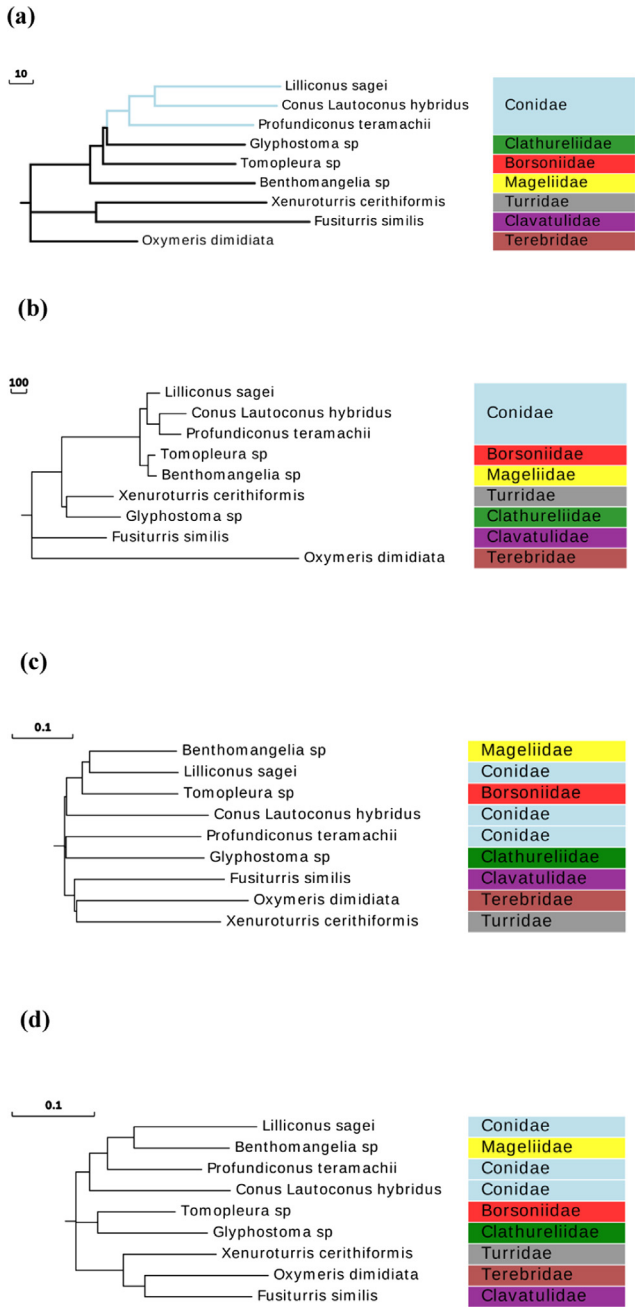


Fig. 6. Phylogenetic trees of *Conoidea* superfamily. (a) The CGR-ENV method. (b) The NV method. (c) The MSSIM method combined CGR. (d) The ClustalW method. The sequences of *Conidae* family don't cluster together in one branch by the NV method, the MSSIM method and the ClustalW method.

Table 1
Time comparison of the three methods.

	Influenza (seconds)	*Bacillus* (seconds)	*Conoidea* (seconds)
CGR-ENV method	0.63	165.72	0.92
NV method	0.02	10.00	0.07
MSSIM method	2.6	526.37	3.2
ClustalW method	70.91	–	892.74
Higher-Oreder Markov model	60.06	2479.44	21.51

method is slower than the NV method, whose complexity is O(kL). The CGR image reflects the frequency of each n-string word, and the ENV uses the distribution of gray values. In contrast, the NV method only measures a single nucleotide's number and position, and cannot capture the information of n-string words ($n > 1$). The Higher-Order Markov model needs to count the frequency of n-string words and (n-1)-string words in each calculation, so it consumes more time.

3.5. Accuracy analysis and phylogenetic trees distance

Moreover, in order to evaluate the accuracy of our alignment-free approach, we calculate the Robinson-Foulds (RF) distances between the phylogenetic trees constructed by our CGR-ENV method, the NV method, the MSSIM method and the Higher-Order Markov model and the phylogenetic trees obtained by using the ClustalW method on *Influenza* and *Conoidea* sequences datasets, respectively. The RF distance between two trees is the number of bipartitions that differ between the trees (Robinson and Foulds, 1981). The result is shown in Fig. 4. For *Influenza* A virus dataset, the results show that the distance between phylogenetic trees obtained using our CGR-ENV method and those generated by the ClustalW method is smaller than the distances between those generated by the NV method and the Higher-Order Markov model. Although the phylogenetic tree of *Influenza* A viruses by the MSSIM method is closer to that by the ClustalW method, there is a sequence of H5N1 misplaced in H1N1 cluster, which is shown in Figure S2 and S3. However, our method can divide the 29 *Influenza* A virus sequences into 5 clusters correctly, which is more accurate than the ClustalW method. So our method performs better than the NV method. For the *Conoidea* dataset, the RF distance for our method is smaller than for the NV method or the MSSIM method. The RF distance for the Higher-Order Markov model is the smallest, but the sequences of *Conidae* family don't cluster together in one branch. So our method is more accurate than the other methods.

4. Discussion

This article proposes a new alignment-free method for genome comparison based on image analysis. We first transform DNA sequences to grayscale images by CGR. Next, we calculate the ENVs of the grayscale images. In the Method section, we prove that ENV makes it possible to recover the grayscale images of the DNA sequences, which preserve more information from the raw data. The performance results we obtain on the three datasets prove that the new proposed CGR-ENV method can be applied to large genomes in order to produce accurate results with high time-efficiency. It provides a new quantitative way of analyzing evolutionary relationships among species in molecular biological study.

The MSA method is time consuming and requires a lot of memory. It relies on a given score matrix, which is the penalty function of mismatch and gap. The results are directly influenced by this score matrix (Dong et al., 2018), so it is artificial. In order to speed up genome comparison, computational and statistical methods to cluster the DNA have been successfully applied in clustering DNA, such as NV method. The NV method only measures a single nucleotide's number and position, but it cannot capture the information of n-string word ($n > 1$). The Higher-Order Markov model only reflects the frequency of an n-string word. But according to (Almeida et al., 2001), CGR images can reflect the frequency of each n-string word, and the ENV also contains position information. So from this perspective, our method performs better than the NV method and the Higher-Order Markov model.

Because the ENVs are in one-to-one correspondence with the CGR images, the ENV method is more reliable than other methods based on CGR images. So our method reflects the similarity between sequences better than the MSSIM method. Meanwhile, the ENV can be used not only in sequence alignment but also in fields related to grayscale image comparison.

5. Funding

This work was supported by Tsinghua University start-up fund, National Natural Science Foundation of China grant (#91746119) and Tsinghua University Education Foundation fund (042202008).

Acknowledgements

Stephen S. -T. Yau and Rong Lucy He conceived the research project; Shaojun Pei and Wenhui Dong carried out the data analysis including figures drawing; Stephen S.-T.Yau provided the idea while Wenhui Dong and Xiuqiong Chen wrote down the proof of the theorem; Stephen S.-T.Yau is the corresponding author and designed the studies; Shaojun Pei and Wenhui Dong wrote the first draft of the paper. All authors participate in writing the final version of the paper. Prof. Stephen Yau is grateful to National Center for Theoretical Sciences (NCTS) for providing excellent research environment while part of this research was done.

Appendix A. Supplementary material

Supplementary Data are available online. The Matlab program of ENV is available at GitHub https://github.com/YaulabTsinghua/Extended-Natural-Vector.

References

Ackermann, H.W., Kropinski, A.M., 2007. Curated list of prokaryote viruses with fully sequenced genomes. Res. Microbiol. 158, 555–566.
Moore, B., Hu, H., Singleton, M., De La Vega, E.M., Reese, M.G., Yandell, M., 2011. Global analysis of disease-related DNA sequence variation in 10 healthy individuals: implications for whole genome-based clinical diagnostics. Genet. Med. 13, 210–217.
Vinga, S., Almeida, J., 2003. Alignment-free comparison – a review. Bioinformatics 19, 513–523.
Ni, H., Qi, D., Mu, H., 2018. Applying MSSIM combined chaos game representation to genome sequences analysis. Genomics 110, 180–190.
Higgins, D.G., Thompson, J.D., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl. Acids Res. 22, 4673–4680.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. 32, 1792–1797.
Notredame, C., Higgins, D., Heringa, J., 2000. T-Coffee: A novel method for multiple sequence alignments. J. Mol. Biol. 302, 205–217.
Katoh, M., Kuma, M., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucl. Acids Res. 30, 3059–3066.
Ludwig, W., Strunk, O., Klugbauer, S., Klugbauer, N., Weizenegger, M., Neumaier, J., 1998. Bacterial phylogeny based on comparative sequence analysis. Electrophoresis 19, 554–568.
Lang, J.M., Darling, A.E., Eisen, J.A., 2013. Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. PloS One 8.
Wang, Z., Wu, M., 2013. A phylum-level bacterial phylogenetic marker database. Mol. Biol. Evol. 30, 1258–1262.
Jeffrey, H.J., 1990. Chaos game representation of gene structure. Nucl. Acids Res. 18, 2163–2170.
Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E., Stanhope, M.J., 2001. Universal trees based on large combined protein sequence datasets. Nat Genet. 28, 281–285.
Almeida, J.S., Carrico, J.A., Maretzek, A., Noble, P.A., Fletcher, M., 2001. Analysis of genomic sequences by Chaos Game Representation. Bioinformatics 17, 429–437.
Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G., Fertil, B., 1999. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol. Biol. Evol. 16, 1391–1399.
Higgins, D.G., Sharp, P.M., 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73, 237–244.
Yang, W.F., Yu, Z.G., Anh, V., 2016. Whole genome/proteome based phylogeny reconstruction for prokaryotes using higher order Markov model and chaos game representation. Mol. Phylogenet. Evol. 96, 102–111.
Lefort, V., Desper, R., Gascuel, O., 2015. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. Mol. Biol. Evol. 32, 2798–2800.
He, Z., Zhang, H., Gao, S., Lercher, M.J., Chen, W.H., Hu, S., 2016. Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. Nucl. Acids Res. 44, W236–W241.
Deng, M., Yu, C., Liang, Q., He, R.L., Yau, S.S.-T., 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. Plos One 6 e17293.
Pei, S., Dong, R., He, R.L., Yau, S.S.-T., 2019. Large-scale genome comparison based on cumulative Fourier power and phase spectra: central moment and covariance vector. Computat. Struct. Biotechnol. J. 17, 982–994.
Webster, R.G., Bean, W.J., Gorman, O.T., Chambers, T.M., Kawaoka, Y., 1992. Evolution and ecology of influenza A viruses. Plos One 56, 359–375.
Sims, G.E., Se-Ran, J., Wu, G.A., Sung-Hou, K., 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc. Natl. Acad. Sci. USA 106, 2677–2682.
Puillandre, N., Samadi, S., Boisselier, M.C., Sysoev, A.V., Kantor, Y.I., Cruaud, C., 2008. Starting to unravel the toxoglossan knot: molecular phylogeny of the turrids (neogastropoda: conoidea). Mol. Phylogenet. Evol. 47, 1122–1134.
Uribe, Juan E., Puillandre, N., Zardoya, R., 2017. Beyond conus: phylogenetic relationships of Conidae based on complete mitochondrial genomes. Proc. Natl. Acad. Sci. USA 107, 142–151.
Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147.
Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: molecular evolutionary genetics analysis Version 7.0 for Bigger Datasets. Mol. Biol. Evol. 33, 1870–1874.
Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.
Dong, R., Zhu, Z., Yin, C., He, R.L., Yau, S.S.-T., 2018. A new method to cluster genomes based on cumulative Fourier power spectrum. Gene 673, 239–250.