Journal of Theoretical Biology 382 (2015) 99–110

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

An improved model for whole genome phylogenetic analysis by Fourier transform

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

[a] Department of Mathematics, Statistics and Computer Science, The University of Illinois at Chicago, Chicago, IL 60607-7045, USA
[b] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China

HIGHLIGHTS

- We propose a 2D numerical representation of a DNA sequence.
- We propose to incorporate nucleotide composition into similarity measure.
- We propose a method to even scale a time series to any lengths.
- We apply the discrete Fourier transform on whole genomes as distance measure.

ARTICLE INFO

Article history:
Received 3 March 2015
Received in revised form
19 June 2015
Accepted 22 June 2015
Available online 4 July 2015

Keywords:
Genome
Similarity distance
Fourier transform
Even scaling
Phylogenetic analysis

ABSTRACT

DNA sequence similarity comparison is one of the major steps in computational phylogenetic studies. The sequence comparison of closely related DNA sequences and genomes is usually performed by multiple sequence alignments (MSA). While the MSA method is accurate for some types of sequences, it may produce incorrect results when DNA sequences undergone rearrangements as in many bacterial and viral genomes. It is also limited by its computational complexity for comparing large volumes of data. Previously, we proposed an alignment-free method that exploits the full information contents of DNA sequences by Discrete Fourier Transform (DFT), but still with some limitations. Here, we present a significantly improved method for the similarity comparison of DNA sequences by DFT. In this method, we map DNA sequences into 2-dimensional (2D) numerical sequences and then apply DFT to transform the 2D numerical sequences into frequency domain. In the 2D mapping, the nucleotide composition of a DNA sequence is a determinant factor and the 2D mapping reduces the nucleotide composition bias in distance measure, and thus improving the similarity measure of DNA sequences. To compare the DFT power spectra of DNA sequences with different lengths, we propose an improved even scaling algorithm to extend shorter DFT power spectra to the longest length of the underlying sequences. After the DFT power spectra are evenly scaled, the spectra are in the same dimensionality of the Fourier frequency space, then the Euclidean distances of full Fourier power spectra of the DNA sequences are used as the dissimilarity metrics. The improved DFT method, with increased computational performance by 2D numerical representation, can be applicable to any DNA sequences of different length ranges. We assess the accuracy of the improved DFT similarity measure in hierarchical clustering of different DNA sequences including simulated and real datasets. The method yields accurate and reliable phylogenetic trees and demonstrates that the improved DFT dissimilarity measure is an efficient and effective similarity measure of DNA sequences. Due to its high efficiency and accuracy, the proposed DFT similarity measure is successfully applied on phylogenetic analysis for individual genes and large whole bacterial genomes.

1. Introduction

DNA sequence comparison is a discipline that has grown enormously in recent years due to the overwhelming burst in sequence data. Discovery of novel biological functions from the *ab initio* analysis of DNA sequence data depends on sequence comparison and classification, thus it has become increasingly important to develop accurate, reliable and efficient similarity measure in sequence analysis. In similarity comparison, phylogenetic analysis provides insights into the hierarchical relationships between genes, genomes and organisms, and thus becomes a

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

http://dx.doi.org/10.1016/j.jtbi.2015.06.033
0022-5193/

fundamental research approach in structure and function analysis of biological sequences (Eisen, 1998). Construction of a phylogenetic tree of DNA sequences has two phases. The first phase is to construct distance matrix from the DNA sequences using either multiple sequence alignment (MSA) or alignment-free methods on DNA sequences. The second phase is to construct the UPGMA or neighbor-joining phylogenetic tree from the distance matrix. The majority of biological sequence comparison methods relies on MSA (Warnow, 2013), however, the sequence alignments become difficult when DNA sequences share low similarities or the sequences are very long because the MSA computational load escalates as an exponential function of the sequence lengths. This problem makes use of MSA for comparing and searching large DNA sequence data infeasible (Edgar and Batzoglou, 2006; Kemena and Notredame, 2009; Chan and Ragan, 2013).

Alignment-free methods, which overcome problems in MSA, have been developed during last decades (Song et al., 2013; Vinga and Almeida, 2003; Patil and McHardy, 2013). The alignment-free methods can be classified into two major categories. The first and widely used approach is based on word frequencies on DNA sequences, in which DNA sequences are converted to feature vectors defined by the frequency of k-mer words of DNA sequence (Blaisdell, 1986, 1989; Sims et al., 2009; Jun et al., 2010). The k-mer words in a DNA sequence are all possible permutations of length k from four nucleotide A, T, C, G. For example, if $k=5$, there are $4^5=1024$ such possible 5-mer fragments. The k-mer method constructs fixed-length feature vectors by counting the frequencies of occurrence of all k-mer in DNA sequences. The other majority of alignment free methods are mostly derived from the k-mer method, for example, k-string composition vector method was proposed for whole proteome prokaryote phylogeny without sequence alignment (Qi et al., 2004). Although the k-mer method has been successfully used in many applications in biological sequence analysis, those distances depend considerably on the parameter k, and how to choose the optimal k depends on varied degrees of divergence in sequence data (Jun et al., 2010). In addition, when k-mer sizes become large, the k-mer method generates very large dimension of frequency vector and has high computational complexity in k-mer string matching. The second category of alignment-free methods are based on genome features including statistical properties of DNA sequences (Kantorovitz et al., 2007; Dai et al., 2013), the chaos game representation (CGR) of genomes (Jeffrey, 1990; Wang et al., 2005), and graph representations (Qi et al., 2011). However, the k-mer based methods and feature based methods are either computationally extensive or lose information within DNA sequences to a certain degree, therefore, these alignment-free methods have limited applications in phylogenetic analysis of whole genomes.

The limitations in MSA and existing alignment-free method underscore the necessity in using full information content of DNA sequences for fast and accurate similarity comparison. An effective solution is to employ Discrete Fourier Transform (DFT), a well established digital processing approach, in DNA similarity comparison. After DNA sequences are converted from symbolic series into numerical series, DFT can be used to analyze the information content within the DNA sequences in frequency domain. The associated Fourier power spectra reflect nucleotide distributions in the sequences, and thus have been used for detecting periodicities of protein-coding genes in genomes (Marhon and Kremer, 2011; Sharma et al., 2004; Marsella et al., 2009; Yin and Yau, 2005, 2007). Previously we presented a novel alignment-free similarity comparison method by Fourier power spectra of DNA sequences with even scaling (Yin et al., 2014). However, that method has a limitation that a DNA sequence cannot be extended to a length of more than twice of its original length. This limitation restricts the general application of the method on highly heterogeneous DNA sequences.

In this paper, we present an improved model for DNA similarity measure based on DFT of DNA sequences. In this model, we propose a new algorithm to map DNA sequences to 2D numerical sequences that incorporates nucleotide composition of the sequences, and therefore similarity distance measure reflects the difference of the nucleotide composition. The new mapping can greatly improve accuracy and significantly increase the computational performance compared with 4D binary indicator representation. In addition, we establish a new even scaling algorithm that can stretch a numerical series to any lengths. This even scaling algorithm can therefore be used to extend the Fourier power spectra of any genomes of any lengths to the same length so that the distance of these genomes can be measured in the same Euclidean space. We assessed the improved DFT method on different DNA datasets in phylogenetic analysis. We demonstrate that the proposed method outperforms the previous method and gives better alignment results than our previous method for different empirical evaluations. Its practical application is expected in genome phylogenetic tree construction and next generation sequencing data studies. We also evaluated the efficiency and accuracy of the proposed DFT method in whole genome phylogenetic analysis, our results demonstrate that a total of 40 full large bacterial genomes can be effectively classified.

2. Methods and algorithms

2.1. Numerical representations of DNA sequences

A DNA molecule consists of four linearly linked nucleotides, adenine (A), thymine (T), cytosine (C), and guanine (G). To apply digital signal processing approaches to a DNA sequence study, the symbolic DNA sequence is mapped into one or more numerical sequences. The commonly used numerical mapping method is Voss 4D binary indicator sequences (Voss, 1992). In the Voss 4D method, a DNA sequence of length N, denoted as $s(0), s(1), \ldots, s(N-1)$, can be decomposed into four binary indicator sequences, $u_A(n), u_T(n), u_C(n)$, and $u_G(n)$, which indicate the presence or absence of four nucleotides, A, T, C, and G at the n-th position, respectively. The Voss 4D binary indicator mapping of a DNA sequence is defined as follows:

$$u_\alpha(n)=\begin{cases}1, & s(n)=\alpha\\ 0 & \text{otherwise}\end{cases} \tag{1}$$

where $\alpha\in\{A,T,C,G\}$, $n=0,1,2,\ldots,N-1$. The four indicator sequences correspond to the distributions of the four nucleotides at each position of the DNA sequence.

To improve the performance of DNA similarity analysis method, here, we propose following 2D numerical representation of a DNA sequence, in which the dimension of the numerical sequences is reduced from 4D to 2D. In 2D numerical representations, we propose that one of the mapping functions β of the four nucleotides A,T,C,G of a DNA sequence can be defined as

$$\beta(A)=[0,-1]',\quad \beta(T)=[-1,0]',\quad \beta(C)=[1,0]',\quad \beta(G)=[0,1]'. \tag{2}$$

The 2D numerical representation of a DNA sequence, $s(0), s(1), \ldots, s(N-1)$, is defined by a 2D matrix v as follows:

$$v(n)=[v_1(n),v_2(n)]'=\beta(\alpha)\quad \text{if } s(n)=\alpha \tag{3}$$

where $\alpha\in A,C,G,T, n=0,1,2,\ldots,N-1$. Thus the computational time of DFT in DNA analysis by the new 2D numerical representation can be reduced to half compared with the Voss 4D representation. In this study, we use the 2D binary representation of a DNA sequence for DFT followed by even scaling in similarity analysis. Table 1 illustrates the 4D Voss representation as u_A, u_T, u_C, u_G and a 2D numerical representation as matrix v of an example DNA sequence.

Table 1
Example of the Voss 4D binary indicator and a 2D numerical mappings of a short DNA sequence.

DNA	T	A	G	C	C	T	G	C	T	G	A	T
u_A	0	1	0	0	0	0	0	0	0	0	1	0
u_T	1	0	0	0	0	1	0	0	1	0	0	1
u_C	0	0	0	1	1	0	0	1	0	0	0	0
u_G	0	0	1	0	0	0	1	0	0	1	0	0
v_1	−1	0	0	1	1	−1	0	1	−1	0	0	−1
v_2	0	−1	1	0	0	0	1	0	0	1	−1	0

2.2. Discrete Fourier transform

Discrete Fourier transform (DFT) is the transformation of observation data in time domain to new values in frequency domain. DFT spectral analysis of DNA sequences may detect latent or hidden periodical signals in the original sequences. It may discover approximate repeats that are difficult to detect by tandem repeat search. Let $X(k)$ be the DFT of time series $x(n)$ of length N, and $X(k)$ is defined as

$$X(k)=\sum_{n=0}^{N-1} x(n)e^{-i(2\pi/N)kn},\quad k=0,1,2,\ldots,N-1 \tag{4}$$

where $i=\sqrt{-1}$. The DFT power spectrum of the signal $x(n)$ at the frequency k is defined as

$$PS_X(k)=\sum |X(k)|^2,\quad k=0,1,2,\ldots,N-1 \tag{5}$$

Let V_j denote DFT of row j of the binary matrix v in Eq. (3) for a DNA sequence, we have the DFT power spectrum of v as

$$PS_V(k)=\sum_{j=1}^{2} |V_j(k)|^2,\quad k=0,1,2,\ldots,N-1 \tag{6}$$

Theorem 2.1 (*Parseval Theorem*). *The total energy contained in a signal x(n) summed across all of time n is equal to the total energy of the Fourier transform X(k) summed across all of its frequency components k. For the discrete Fourier transform (DFT), the relation is*

$$\sum_{n=0}^{N-1} |x(n)|^2=\frac{1}{N}\sum_{k=0}^{N-1} |X(k)|^2$$

where X[k] is the DFT of x[n], both of length N.

The Parseval's theorem on Fourier transforms implies the equivalence in the energy levels of signal in frequency domain and time domain. The Fourier transform preserves the Euclidean distance between two signals (Faloutsos et al., 1994; Agrawal et al., 1993). Therefore, the Fourier transform gives a unique representation of the original underlying signal in frequency domain, in which the numerical vector in the frequency domain contains all the information about signal in the time domain. We can infer information content in DNA sequences from the distribution of Fourier power spectra of the sequences, and use the Euclidean distances of the Fourier power spectra of DNA sequences as the similarity measure.

One may question that there are 24 mappings between the four nuleotides and the 2D numerical representations, the choice of one over the other mapping for DNA sequences may produce different DFT distances. We have the following theorem and algorithm to address this question.

Theorem 2.2. *For the 24 mappings between the four nuleotides (A,T, C,G) and the four 2D numerical representations, there are three distinct power spectra, each of the unique spectrum corresponds to eight 2D mappings. The three distinct 2D mappings, named as 2D-AT, 2D-AC, 2D-AG, are in Table 2. The corresponding eight 2D mappings are in Tables A1, A2 and A3 in Supplementary Materials.*

Table 2
Three unique 2D mappings of DNA sequences.

Base	2D-AT	2D-AC	2D-AG
$A=$	0	0	0
	−1	−1	−1
$T=$	0	−1	−1
	1	0	0
$C=$	−1	0	1
	0	1	0
$G=$	1	1	0
	0	0	1

The proof of this theorem is straightforward. A mapping in one of the three groups 2D-AT, 2D-AC, 2D-AG (Tables A1, A2 and A3 in the Supplementary Materials) is either a rotation or a reflection of the other in the same group, because DFT is linear and orthogonal transform, Parseval's theorem on Fourier transform indicates the equivalence of all the eight 2D mappings within one group.

It is now well-established that nucleotide composition is more similar from closely related organisms than for distantly related ones. The nucleotide composition has a bias on C+G contents, while A vs T, or C vs G content is similar across different organisms. The genomic percentage of G+C content is highly variable among prokaryotes and other unicellular organisms (Hildebrand et al., 2010). Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions (Foster and Hickey, 1999; Mrázek, 2009). Therefore, in the three unique 2D mappings of DNA sequences (Table 1), we exclude the 2D-AT mapping for DNA sequence because 2D-AT mapping distributes the A+T, or the C+G content, into the same rows and thus makes 1/−1 and 0 uneven on the two rows. The choice of 2D-AC or 2D-AG mappings depends on which content (A+C or A+G) is close to 50% of base compositions in a DNA sequence so that the 1/−1 and 0 can be evenly distributed on the two rows of the 2D mapping. Thus, the correspondence between a DNA sequence and its 2D mapping by the mapping algorithm is one-to-one and no parameter is required in the corresponding DFT distance measure. These considerations have been validated on real genes and genomes of different organisms in this study. The detailed method of the 2D mapping of a pair of DNA sequences based on nucleotide compositions is described in Algorithm 1.

Algorithm 1. Getting the 2D mapping of a pair of DNA sequences based on nucleotide compositions.

Input: DNA sequences: SEQ1(length N1), SEQ2(length N2)
Output: 2D numerical mapping for SEQ1 and SEQ2
Step:
1. Compute A+C, A+G of SEQ1 and SEQ2 together.
2. Compute values:

$$R_{AC}=\left|\frac{\mathrm{A+C}}{\mathrm{N1+N2}}-\frac{1}{2}\right|,\quad R_{AG}=\left|\frac{\mathrm{A+G}}{\mathrm{N1+N2}}-\frac{1}{2}\right|$$

3. **if** $R_{AC}<R_{AG}$ **then return** 2D-AC
else return 2D-AG

Comparing the 4D binary and 2D numerical representations, the advantage of the 4D binary indicator representation of a DNA sequence is that it does not predefine any mathematical relationship among the symbols and only indicates the frequencies of the symbols. Thus it is widely utilized in detecting symbol distributions and periodicity features of a sequence. However, using the

4D indicator representation, two different DNA sequences by 4D representation may have the same power spectra. For example, let u_{1A}, u_{1T}, u_{1C}, and u_{1G} be the four indicator sequences of the DNA sequence $S_1 = ATCGAA$, and let u_{2A}, u_{2T}, u_{2C}, and u_{2G} denote the four indicator sequences of the DNA sequence $S_2 = GCTAGG$, $u_{1A} = u_{2G}$, $u_{1T} = u_{2C}$, $u_{1C} = u_{2T}$, and $u_{1G} = u_{2A}$, the power spectra of these two different DNA sequences are the same. The proposed 2D numerical representation addresses the problem of uncertainness in the Fourier power spectra of DNA sequences.

2.3. Even scaling of Fourier power spectrum of different lengths

From the definition of Fourier power spectrum, DNA sequences of different lengths have power spectra of different lengths and thus the power spectra cannot be used as a direct comparison of DNA sequences. In the literature, a solution is to use partial spectra from a few beginning frequencies (Wu et al., 2000; Wang et al., 2013; Rafiei and Mendelzon, 1998), but this approach loses information for sequence comparison. To overcome the above problem, we propose here the following even scaling method to scale the DFT power spectra of different lengths into the same length. We take one or two consecutive data elements in the shorter data series to evenly stretch the short data series to a new length. In detail, let T_n denote the original power spectrum of length n and T_m denote the extended power spectrum of length m from even scaling of T_n and $m > n$. The symbol $\lfloor \cdots \rfloor$ denotes the floor function on non-integers. The even scaling operation on the original power spectrum T_n to T_m is defined as follows:

$$T_m(k) = \begin{cases} T_n(Q) & \text{if } Q \in Z^+ \\ T_n(R) + (Q-R)(T_n(R+1) - T_n(R)) & \text{if } Q \notin Z^+ \\ \text{where } Q = \dfrac{kn}{m}, \quad R = \left\lfloor \dfrac{kn}{m} \right\rfloor & \end{cases} \tag{7}$$

The even scaling method is assessed by statistical central moments and complexity evaluation. The complexity estimate (CE) of a time series T_n of length n is normalized from Batista et al. (2011) and is defined as

$$CE = \sqrt{\frac{\sum_{k=1}^{n-1} (T_n(k) - T_{-n}(k+1))^2}{\sum_{k=1}^{n} (T_n(k))^2}} \tag{8}$$

It is worthy to note that the proposed even scaling algorithm can scale up to any length. This property makes the even scaling method flexible in different application perspectives. The even scaling method is described in detail in Algorithm 2.

Algorithm 2. Even scaling a number series T_n.

Input: T_n of length n, new length m, $m > n$
Output: T_m of length m
$T_m(1) = T_n(1)$
for $k \leftarrow 2$ ***to*** m **do**
  $Q = \frac{kn}{m}$
  $R = \lfloor \frac{kn}{m} \rfloor$
  if $R == 0$ **then**
    $R \leftarrow 1$
  end
  if $Q \in Z^+$ **then**
    $T_m(k) = T_n(Q)$
  else
    $T_m(k) = T_n(R) + (Q - R) * (T_n(R+1) - T_n(R))$
  end
end
return T_m

2.4. Algorithm for pairwise Euclidean distances of DNA sequences in Fourier frequency domain

The most common distance measure for time series is the Euclidean distance, which is the optimal distance measure for estimation if signals corrupted by additive Gaussian noise (Agrawal et al., 1993). The *Euclidean metric* on $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is defined by the function d:

$$d((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \tag{9}$$

A true distance metric for DNA sequences shall satisfy the triangle inequality of metric space. It specifies that direct path between two sequences cannot be longer than a less-direct path involving other intermediate sequence. If a distance metric that does not conform to this relation are nonmetric and is internally inconsistent (Wheeler, 1993). After even scaling the DFT spectra, we measured the Euclidean distances of DNA sequences using the full DFT power spectra of the DNA sequences. Since we embed all the DNA sequence information via their full power spectrum into the same Euclidean space, the induced distance metric we propose here is true metric.

It is worth mentioning that the distance measure in Fourier frequency domain in this study excludes the zero-th term in the power spectrum. Because the zero-th power spectrum is just the sum of data, its values usually are much larger than the rest of the power spectrum. If the zero-th power spectrum value is included in the Euclidean distance calculation, the accuracy of the similarity measure of DNA sequences is reduced.

The detailed method to compute the Euclidean distances of DNA sequences in Fourier frequency domain of two DNA sequences is described in Algorithm 3.

Algorithm 3. Computing the Euclidean distances of DNA sequences in Fourier frequency domain.

Input: DNA sequences SEQ1(length N1) and SEQ2(length N2), common length M, and M > N1, M > N2
Output: Euclidean distance of SEQ1 and SEQ2
Step:
1. Get 2D mapping method for SEQ1 and SEQ2 based on nucleotide composition (Algorithm 1).
2. Convert SEQ1 and SEQ2 to 2D numerical sequence based on the 2D mapping method.
3. Compute the corresponding Fourier power spectrum PS1 and PS2 from the converted 2D numerical vectors (Eq. (6)).
4. Even scale PS1 from length N1 to length M, named PSM1 (Algorithm 2).
5. Even scale PS2 from length N2 to length M, named PSM2.
6. Compute the Euclidean distance d(P1M,P2M) in an M-dimensional space (Eq. (9)).

DNA similarity analysis is performed using UPGMA (Unweighted Pair Group Method with Arithmetic Mean) hierarchical clustering method from a pairwise distance matrix (Sourdis and Krimbas, 1987). The UPGMA method builds the phylogenetic tree bottom up from its leaves for the given set of DNA sequences. It constructs each DNA sequence to form a cluster first, then groups two smaller clusters of nodes recursively until there is only one phylogenetic tree that contains all the DNA sequences. The resulting UPGMA tree reflects the structure and relationship of the sequences presented in the distance matrix.

2.5. Implementations and data

For comparison purpose, we used the following similarity measures of DNA sequences: (1) The proposed DFT method with even scaling that is implemented in MATLAB R2011b. (2) MSA method with the Jukes–Cantor genetic distance measure. The Jukes–Cantor genetic distance is the maximum likelihood estimate of the number of substitution that occurred per site over the course of one sequence evolving from another. The pairwise sequence alignment was performed using MATLAB R2011b bioinformatics toolbox. (3) Alignment-free *k*-mer method. The pairwise distance of the *k*-mer frequency vectors of different DNA sequences was measured by the Euclidean distance. The *k*-mer method used in this study is from the MATLAB NACS toolbox v4.1 (Vinga and Almeida, 2003).

All sequence data were obtained from GenBank on the NCBI and are listed in the Supplementary Materials. The methods and algorithms in this study were implemented in MATLAB language and are available from the following site: http://www.mathworks.com/matlabcentral/fileexchange

3. Results and discussion

3.1. Even scaling of Fourier power spectra of DNA sequences

We evaluated the effectiveness of the even scaling method by comparing the feature consistency of DFT power spectra of DNA sequences before and after scaling. The even-scaling method was applied to the DFT power spectra of DNA sequences to evenly extend the spectra to different lengths. Fig. 1(a) and (b) is the DFT power spectrum of an exon segment of *Bubo bubo voucher* NHMO-BC120 cytochrome oxidase subunit 1 gene (GenBank ID: GU571285, 360 bp), and it is evenly scaled to 750 bp, respectively. Fig. 1 shows that the scaled data preserve features from the original data, indicating feature consistency among the original and its scaled data. Table 3 is the statistical summary of the original power spectrum and its stretching and shrinking values by even scaling. The means of original and scaled sequences are close to identical and the variances of original and scaled sequences are close. The 3rd central moment (skewness) and 4th central moment (kurtosis) of original are at similar level. The proposed even scaling method overcomes the limitation which requires the shortest length of a DNA more than one-half of the maximum length of the DNA compared. In this even scaling method, the DFT spectra of the short length DNA sequences can be evenly scaled to any lengths.

3.2. Representation of DNA sequences by 2D numerical sequences

To reduce time in computing power spectra of DNA sequences, we represent DNA sequences by 2D numerical vectors, instead of 4D binary indicator sequences. Because the most time spent in computing the similarity distance is on Fourier transform of the numerical vectors that DNA sequences are mapped, the dimension reduction in DNA representation can reduce the computational time to half. The effectiveness of the proposed DFT distance measurement in 2D was tested on single gene. The test gene is the NADH dehydrogenase subunit 4 genes of 12 species of four different groups of primates. The data source consists of four species of old-world monkeys (*Macaca fascicular, Macaca fuscata, Macaca sylvanus,* and *Macaca mulatta*), one species of new-world monkeys (*Saimiri scirueus*), two species of prosimians (*Lemur catta* and *Tarsisus syrichta*), and five hominoid species (*Human, Chimpanzee, Gorilla, Orangutan* and *Hylobates*) (Qi et al., 2011). Fig. 2 (a) and (b) is phylogenetic trees of 12 primate species by DFT with 2D numerical mapping and MSA, respectively. Based on nucleotide

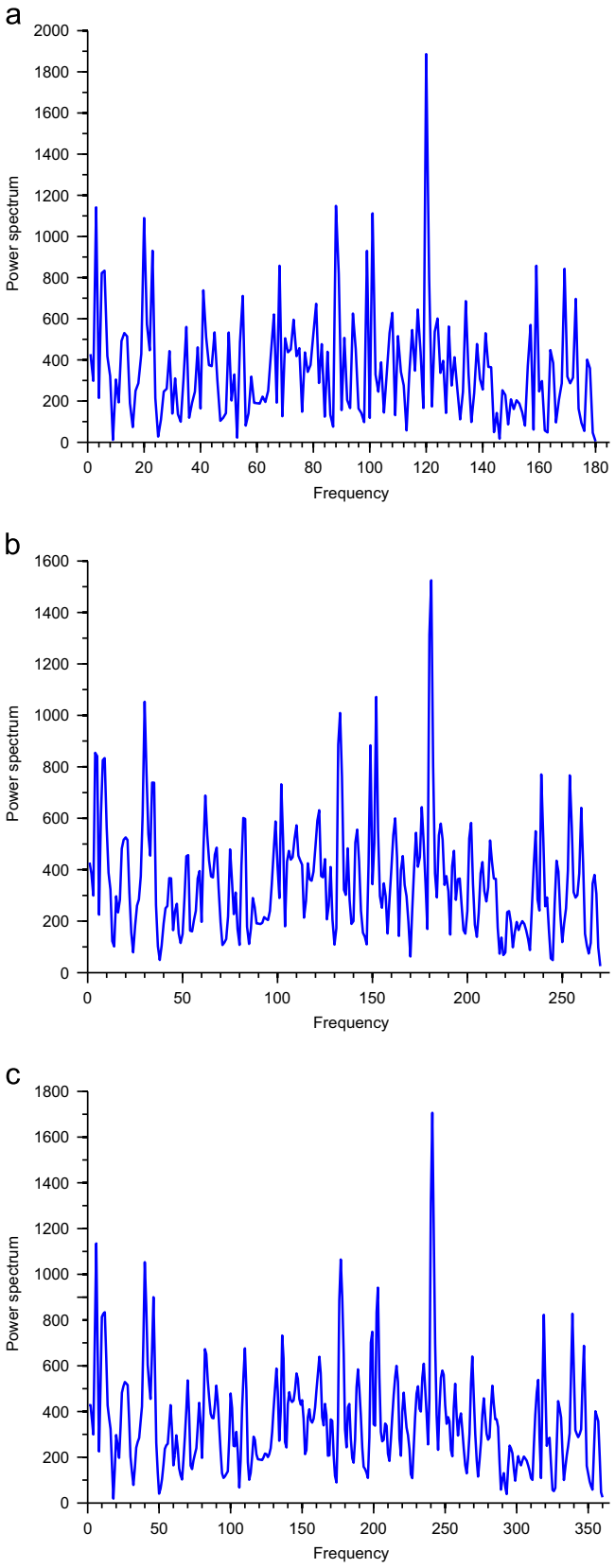

Fig. 1. (a) Fourier power spectrum of Bubo bubo voucher NHMO-BC120 cytochrome oxidase subunit 1 (COI) (360 bp), and even scaled Fourier power spectrum to (b) 540 bp, (c) 720 bp. Because Fourier power spectra of real number series are symmetric, the plots only show the first half of the spectra.

Table 3
Statistical summary of even scaling method.

Length	Mean	Variance	3rd moment	4th moment	CE
360	357.7437	9.2136×10^4	1.3089×10^8	3.4761×10^{11}	0.8924
540	357.7632	6.4106×10^4	7.5916×10^7	1.7310×10^{11}	0.5298
750	357.7694	6.2300×10^4	6.6071×10^7	1.3588×10^{11}	0.3961

composition the 2D mapping is 2D-AG for the NADH dehydrogenase subunit 4 genes of all the 12 species. The result shows that the DFT distance with the 2D mapping generates a phylogenetic tree that is very similar to the tree from MSA. The only difference between the two phylogenetic trees from Fig. 2(a) and (b) is the position of Orangutan due to the fact that there is a deletion mutation in Orangutan (Yin et al., 2014). This deletion cannot be identified by MSA method, and is in agreement with our previous study (Yin et al., 2014). We also compared the accuracy of phylogenetic trees from the DFT distances by the 2D numerical (Fig. 2(a)) and 4D binary mapping (Fig. 2(c)). The phylogenetic tree from 4D mapping has some degree of difference compared with the one from MSA, but the phylogenetic tree structures of these two representations are almost identical. These results suggest that the 2D binary sequence representation proposed in this study can achieve the same accuracy as MSA, while the 4D binary indicator sequences show similar but different tree structures compared with MSA. Furthermore, the 2D representation only uses half computational time compared with the 4D representation. The 2D representation significantly improves the computational performance and is useful, especially, in phylogenetic analysis for full large bacterial genomes.

3.3. *Simulation of construction of phylogenetic trees on different DNA mutations*

A similarity measurement between two DNA sequences shall account for differences in sequences due to insertions, deletions and substitutions of bases in the sequences. These differences are quantified as edit distance. We evaluated the accuracy of the proposed DFT similarity measure using a series of deletion mutations of an intron sequence from 3' end. The deletion size is from 1 bp to 100 bp. We measured DFT similarity distance between the deletion mutants and the original sequence. Fig. 3(a) is the correlation between the deletion lengths and the DFT similarity distances between the corresponding deletion mutants and original sequence. The result in Fig. 3(a) shows a sound linear relationship of the DFT distances and the deletion lengths, demonstrating a robust and reliable behavior of the DFT distance metric in measuring the different lengths of sequences.

The accuracy of the similarity distance metric was also assessed using a series of point mutations (substitutions) in DNA sequences. An intron sequence introduced many different point mutations randomly and the derived mutated sequences were used in the test. We measured the sequence distance between the mutants and the original sequence by the proposed DFT method. Fig. 3(b) is the correlation between the amount of point mutations and the distance between the corresponding point mutants and original sequence. The result in Fig. 3(b) shows sound linear relationship of DFT distances and the amount of point mutations. This result demonstrates the accuracy of the DFT distance metric on the difference of nucleotide mutations on the same length DNA sequences. The above results indicate an equivalency in DFT distance and edit distance in DNA sequences.

One may argue that if we represent four nucleotides, A, T, C, and G, by one dimensional numbers, for example, A=1, T=2, C=3, and G=4, or A=1+i, T=1−i, C=−1+i, and G=−1−i, the computational time for similarity comparison can be better than 2-D representations. But this is not the case because one dimensional representations apply arbitrary mathematical operations or weights on the four nucleotides. For example, that A=1 and T=2 means A is

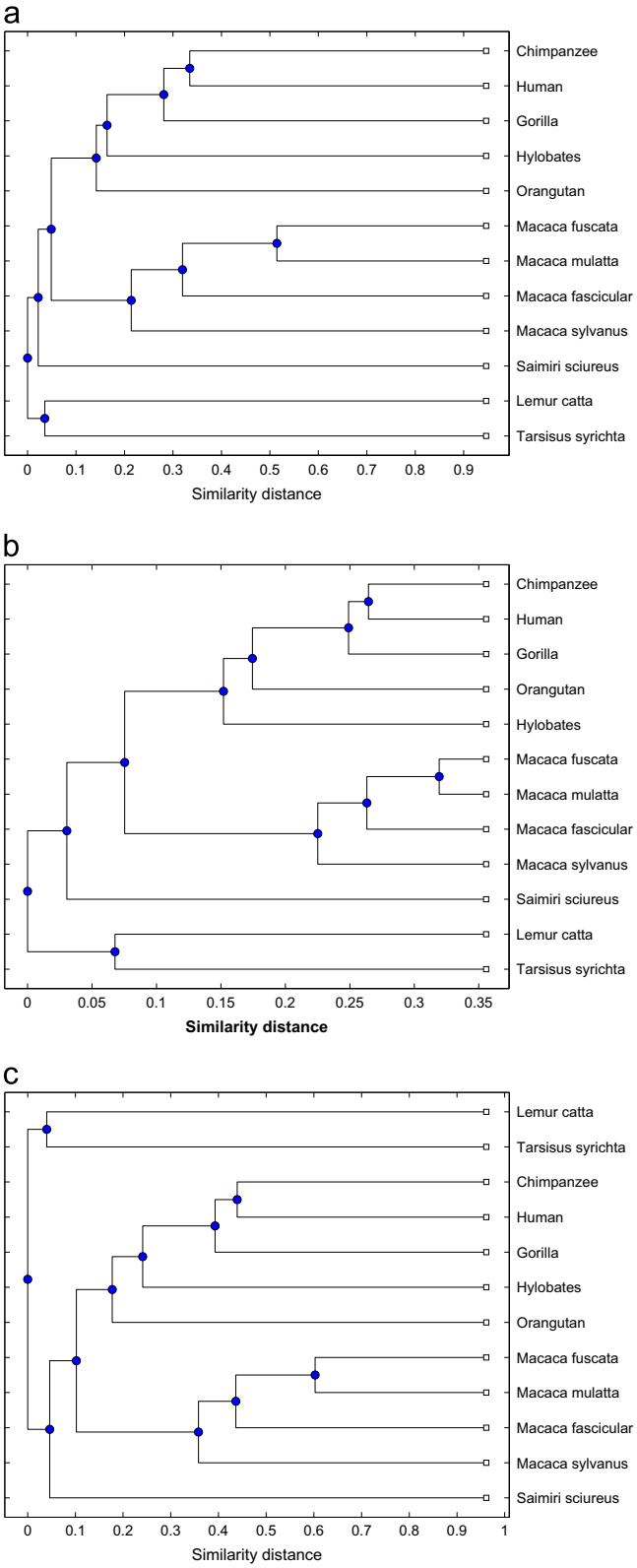

Fig. 2. Phylogenetic tree of 12 primate species on NADH dehydrogenase subunit 4 gene. (a) By the DFT distances of DNA sequences with the 2D numerical mapping, (b) by MSA, (c) by the DFT distances of DNA sequences with the 4D binary indicator mapping.

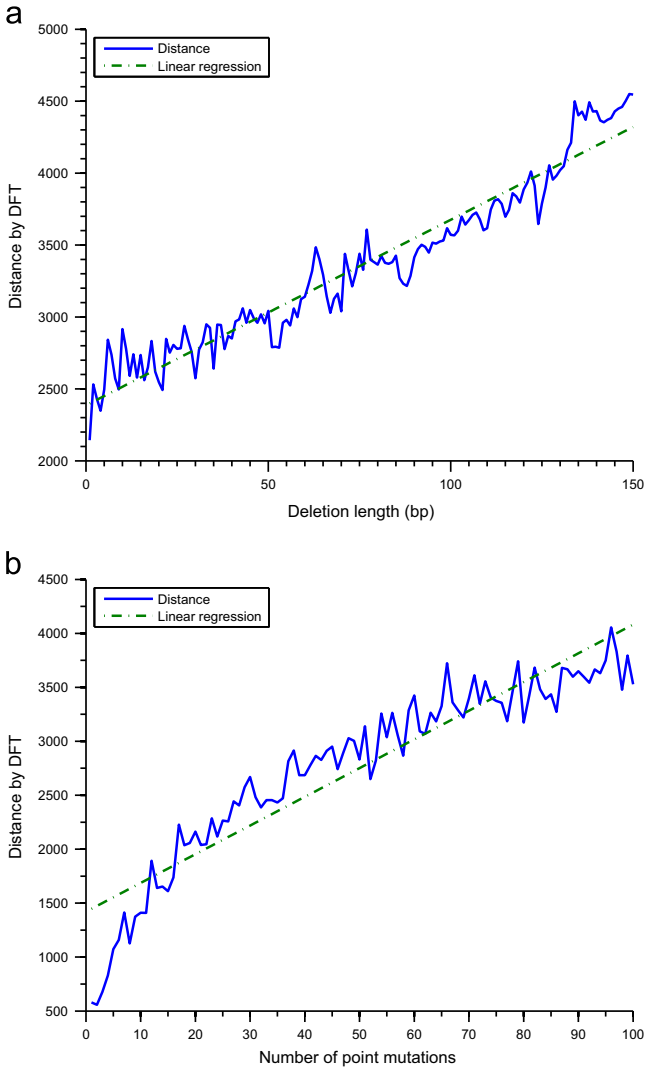

Fig. 3. (a) Correlation of the DFT distance and the lengths of deletion mutants of DNA sequences. (b) Correlation between DFT distance and the number of point mutations of DNA sequences.

smaller than T, but in DNA sequences, A and T shall have equal weights for similarity analysis. For different one dimensional representations, we also tested the correlations between the edit distances and the DFT distances of different deletion or site mutations, but the results do not show linear correlations (data not shown). This study confirms that 2D representation is more accurate than one dimensional representations for similarity analysis of DNA sequences.

To verify if the similarity distance can be used for hierarchical clustering DNA sequences, we generated different mutations in DNA sequences and constructed phylogenetic trees from the pairwise DFT distances of these mutants. We used an intron sequence as base sequence (GeneBank ID: AAG00896, 350 bp) and generated two new sequences A and B from the intron sequence using point mutations. 10% of mutations were introduced into A and B. We then similarly evolved A and B into different mutants by four different mutations (substitutions, deletion, insertion, and transposition). Table 4 is the description on the simulated DNA sequences with different mutations. UPGMA phylogenetic trees of the mutations are built from the distance matrices using the proposed DFT similarity method, alignment-free *k*-mer words method, and pairwise sequence alignment, as shown in Fig. 4(a), (b) and (c), respectively. For the different substitution mutations of the sequence A, Fig. 4(a)–(c) shows that the three methods can correctly classify and cluster them with correct tree topology. All the three methods create the same tree topology corresponding to the numbers of substitution mutations in the DNA sequences. This result indicates that the proposed DFT measure, the *k*-mer and MSA methods have the same discrimination power for measuring substitution mutations. For deletion and insertion mutations of the sequence B, Fig. 4(a)–(c) shows the topological differences in DFT based measure and *k*-mer method and MSA method. Deletion and insertion are two serious mutations and most deletion and insertion mutations may impact significant changes on phenotypes. Fig. 4(a) shows that DFT method can clearly separate the 5NT substitutions from 5 bp deletion or insertion mutations, but *k*-mer and MSA method cannot distinguish these deletion/insertion mutations from substitutions, mixing them in same branches (Fig. 4(b) and (c)). For transposition mutations, Fig. 4(a)–(c) also shows the topological differences in DFT based measure and *k*-mer method and MSA method. Transposition and insertion/deletion are different from substitutions because they cause serious phenotype changes and may be

Table 4
DNA sequence mutation description in simulation tests.

Sequence name	Description
A/original	Generated from AAG00896 (GeneBank ID, 350 bp)
A/substitution/2 NTs/1	2 random nucleotide substitutions in A
A/substitution/2 NTs/2	2 random nucleotide substitutions in A
A/substitution/5 NTs/1	5 random nucleotide substitutions in A
A/substitution/5 NTs/2	5 random nucleotide substitutions in A
A/substitution/10 NTs/1	10 random nucleotide substitutions in A
A/substitution/10 NTs/2	10 random nucleotide substitutions in A
B/original	Generated from AAG00896 (GeneBank ID, 350 bp)
B/substitution/2 NTs/1	2 random nucleotide substitutions in B
B/substitution/2 NTs/2	2 random nucleotide substitutions in B
B/substitution/5 NTs/1	5 random nucleotide substitutions in B
B/substitution/5 NTs/2	5 random nucleotide substitutions in B
B/substitution/10 NTs/1	10 random nucleotide substitutions in B
B/substitution/10 NTs/2	10 random substitution mutations in B
B/deletion/5 bp/51:55	5 bp deletion from positions 51:55 in B
B/deletion/5 bp/101:105	5 bp deletion from positions 101:105 in B
B/insertion/5 bp/51:55	5 bp insertion at position 51 in B
B/insertion/5 bp/101:105	5 bp insertion at position 101 in B
B/transposition/5 bp/50– > 150	5 bp transposition from position 50 to 150 in B
B/transposition/5 bp/50– > 250	5 bp transposition from position 50 to 250 in B

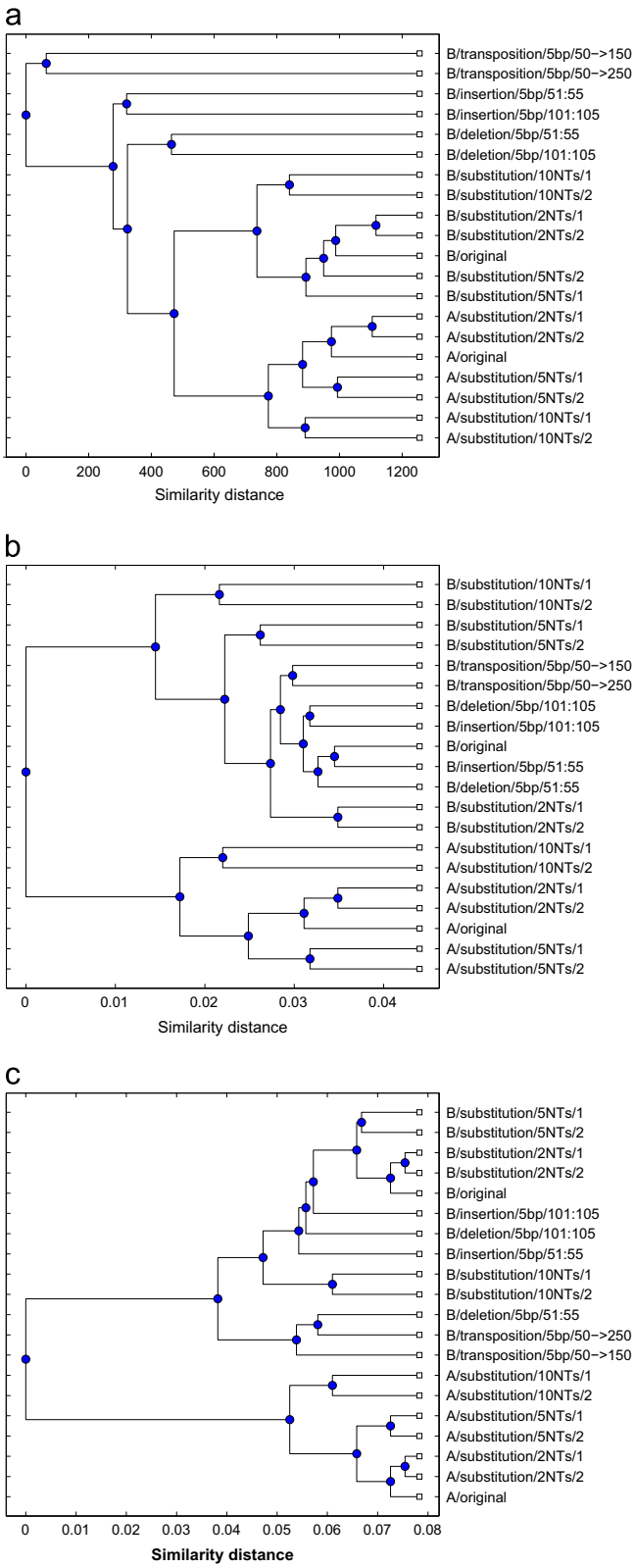

Fig. 4. Clustering analysis of different mutations by phylogenetic trees of simulated DNA sequences in Table 4. (a) The DFT distance, (b) the *k*-mer words, (c) pairwise sequence alignment.

detrimental mutations in hosts. Fig. 4(a) shows DFT method can clearly separate the 5 bp transposition from both substitutions and insertion/deletion mutations, but *k*-mer and MSA method cannot separate transposition mutant from substitutions, mixing them in same branches as shown in Fig. 4(b) and (c). The phylogenetic trees using the proposed DFT distance showed the highest congruences with conventional taxonomic groupings, leading to reliable results in hierarchical clustering of DNA sequences. In some cases, the DFT based phylogenetic trees demonstrate a better identification of different mutations in hierarchical tree and improves computational speed over the *k*-mer method and MSA. The results from 2D representation with new even scaling are in agreement with what we had before using 4D representation (Yin et al., 2014). These results show that the proposed DFT similarity measure can achieve same accuracy and reduce computational time to half compared with our previous method.

3.4. Phylogenetic analysis on individual genes

The utility of the proposed DFT distance measurement was tested on *Influenza* A viruses individual gene level. *Influenza* A viruses cause influenza in birds and domestic poultry and can be occasionally transmitted to human and give rise to human influenza pandemics such as pandemic H1N1/2009 (Vijaykrishna et al., 2010). *Influenza* A viruses are negative-sense, single-stranded, segmented RNA viruses, and can be classified in different subtypes by an H number for the type of hemagglutinin and an N number for the type of neuraminidase. There are 18 different H antigens (H1–H18) and 11 different N antigens (N1–N11). For example, the H5N1 virus designates an *Influenza* A subtype that has a type 5 hemagglutinin (H) protein and a type 1 neuraminidase (N) protein. Using *Influenza* A virus neuraminidase (NA) gene, we constructed phylogenetic tree based on pairwise DFT distance of the segment 6 neuraminidase (NA) gene of different *Influenza* A strains. Figs. 5 and 6 are the phylogenetic trees of *Influenza A* virus constructed by the proposed DFT method and MSA, respectively. The 2D mapping based on nucleotide composition of the virus neuraminidase (NA) genes is 2D-AC. The results in Figs. 5 and 6 indicate that both MSA and DFT trees show correct grouping of different virus subtypes H1N1, H5N1, H3N8, H3N2, H7N3, H11N9, and H7N9. But the phylogenetic tree from DFT distance shows clear branch difference than the phylogenetic tree from the Jukes–Cartor distance in MSA. For example, the virus of highly homologous sequences such as A/Illinois H1N1 viruses and Alaska H7N3 viruses cannot be separated by sequence alignment measured by Jukes–Cartor distance, but they are clearly separated with correct hierarchical relationship in the tree of DFT method. Another example in Fig. 5 is that the H7N9 virus mutants in China 2013 can only be clearly separated in the tree of DFT method. The hierarchical relationship among the H7N9 virus mutants in China is in agreement with the geographic distribution of the virus and the epidemiological investigation from previous findings (Xiong et al., 2013). These results demonstrate the superiority of the proposed DFT method on existing sequence alignment methods due to the fact that the DFT distance calculation is based on all the sequence information and does not lose sequence information after Fourier transform. From disease perspective, the phylogenetic tree using the proposed DFT distance may accurately and rapidly classify and trace the viruses, providing an effective tool for virus surveillance.

3.5. Construction of phylogenetic trees on whole genomes

The comparison of whole genomes has become a very powerful mean for inferring evolutionary relationships because sometimes there is no signature gene available for new genomes. We evaluated and applied the proposed DFT similarity measure on hierarchical clustering genomes, which contain different genes and non-coding regions. Phylogenetic analysis on mitochondrial genes has played an increasingly important role in confirming

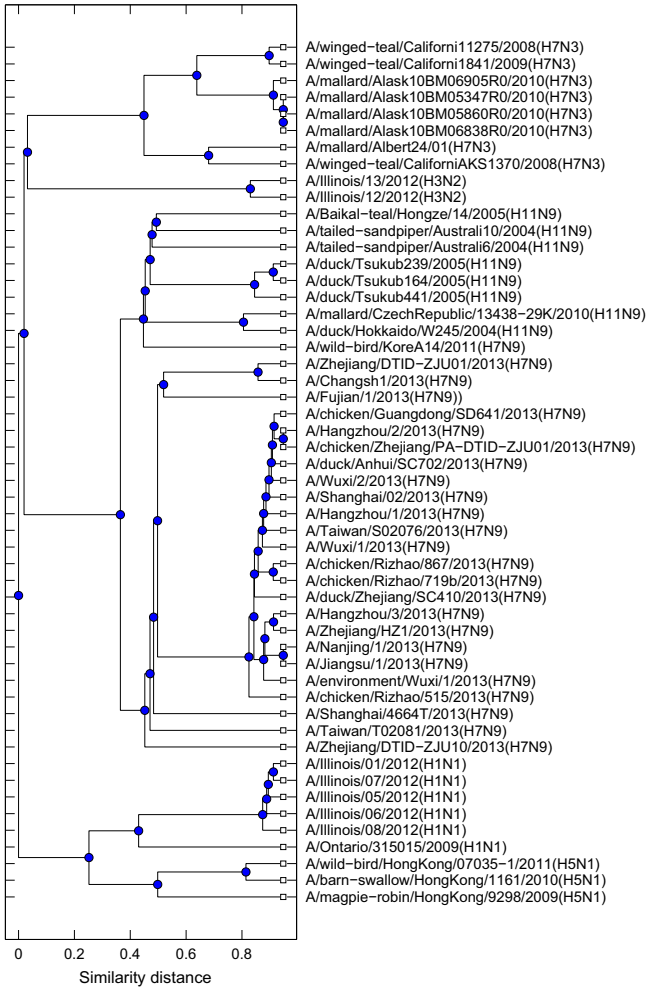

Fig. 5. Phylogenetic tree of *Influenza* A viruses by the DFT distances of DNA sequences with the 2D numerical mapping.

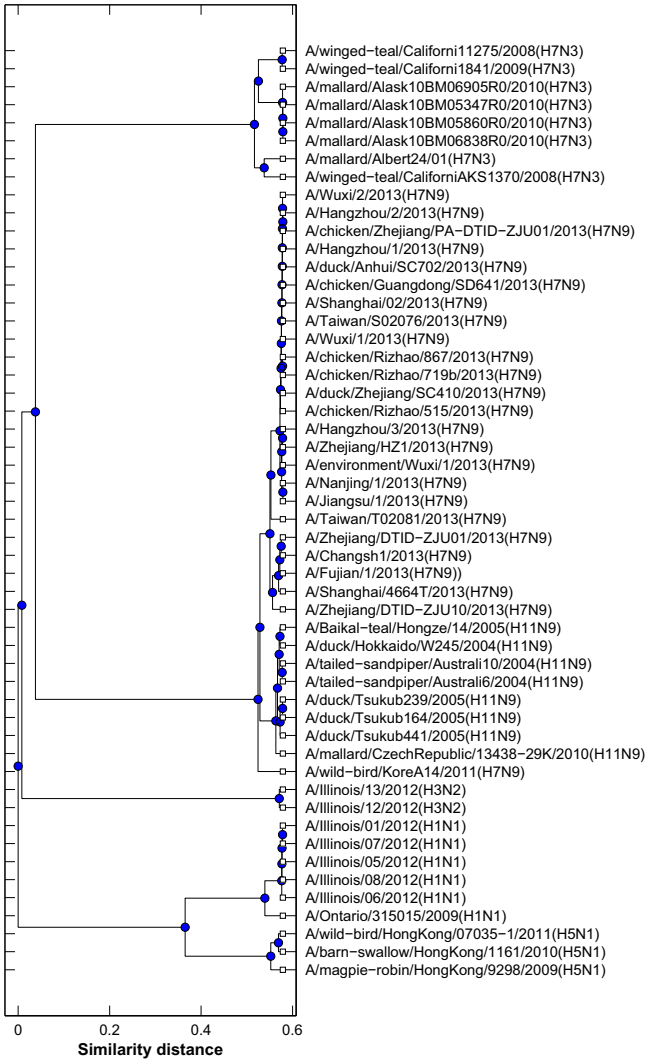

Fig. 6. Phylogenetic tree of *Influenza* A viruses using MSA by Jukes–Cartor distance.

existing or establishing sometimes radically different mammalian groupings and taxonomy (Boore, 1999). For example, previously Tobe et al. (2010) used cytochrome b and cytochrome oxidase subunit I (COI) mitochondrial genes to reconstruct mammalian phylogenies and accurately reconstructs their phylogeny. We constructed phylogenetic tree of 70 mammalian whole mitochondrial genomes using the DFT distance. The 2D mapping based on nucleotide composition of the COI genes is 2D-AG. Fig. 7 shows the taxon relationships inferred from the phylogenetic tree are in agreement with morphological analyses at order, family and generic levels. This result confirms the effectiveness of the proposed DFT similarity measure on DNA similarity analysis.

We assessed the efficiency of the DFT distance in phylogenetic analysis of long whole genomes, which are difficult to compute by genome-wide alignments. We used a total of 40 bacterial organisms with genome sizes from 910k to 5.5M bp. Fig. 8 shows that the use of the DFT distance methods by 2D mapping leads to a reliable phylogenetic tree. All the bacterium including *Bacillus*, *Borrelia*, *Clostridium*, *Desulfovibrio*, *Escherichia*, *Rhodobacter*, *Salmonella*, *Shigell*, *Staphylococcus*, *Sulfolobus*, *Thermoplasma*, and *Yersinia* are clearly classified. It shall be noted that *Escherichia* are very divergent strains. Such large differences in genome size can be mitigated by the proposed even scaling method. Because the nucleotide compositions of these bacterial genomes vary significantly, the 2D mapping based on nucleotide composition of the genomes can be both 2D-AC and 2D-AG, and mitigate the impact of the nucleotide composition bias on distance measure of the whole genomes. Fig. 8 shows that *Escherichia coli* O15:H7 pathogen is divergent from non-pathogen strains K-12 and is close to *Staphylococcus* from the phylogenetic analysis, suggesting horizontal gene transfer between *Staphylococcus* and *Escherichia coli* O157H7, and the horizontal gene transfer was also revealed in previous literatures (Brisson-Noel et al., 1988; Mazodier and Davies, 1991). We compared the 2D mapping and 4D indicator mapping in phylogenetic analysis of the same bacterial genomes. Phylogenetic tree from 4D mapping is shown in Fig. 9. Comparison of Figs. 8 and 9, the 4D representation does not show clear hierarchy relationship from *Yersina* to *Escherichia*, but 2D mapping can show the relationship of this group of bacteria. This result shows that 2D mapping is better than 4D indicator mapping. The reason that 2D mapping can differ from close related bacterial genomes is that the 2D mapping employ nucleotide composition of the genomes. In addition, as evidenced in this study, accurate phylogenetic analysis of *Escherichia coli* O15:H7 provides critical insights into this pathogen's evolutionary patterns such as horizontal gene transfers.

Although our method shows promising results for phylogenetic analysis of whole bacterial genomes, due to the high divergences of bacterial genomes, the results on whole genome analysis from our model are not perfect. For example, we found that in the phylogenetic tree, *Escherichia*/K-12MC4100 is close to *Clostridium perfringens* ATCC 13124, and *Thermoplasma volcanium* GSS1 is close

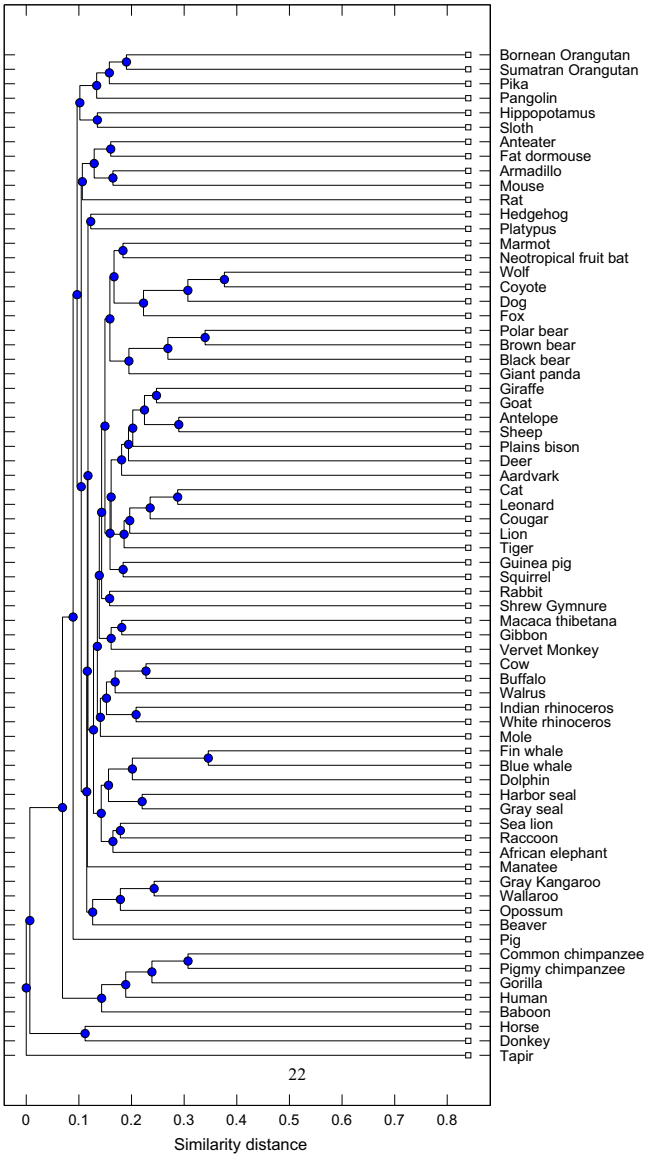

Fig. 7. Phylogenetic tree of SARS viruses by the DFT distances of DNA sequences with 2D mapping.

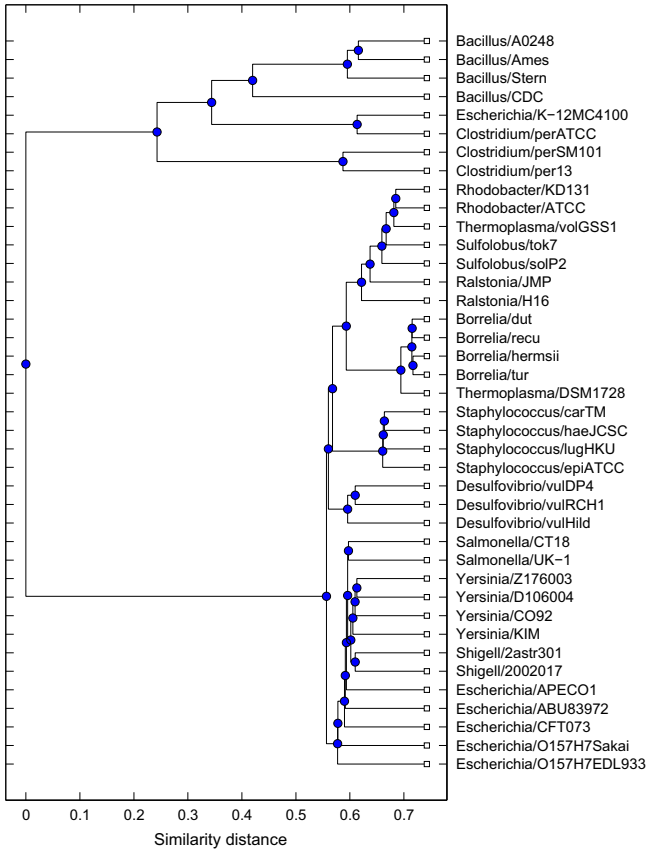

Fig. 8. Phylogenetic tree of 70 mammalian mitochondrial genomes by the DFT distances of DNA sequences with the 2D numerical mapping.

to *Sulfolobus tokodaii str.* 7 and *Rhodobacter sphaeroides* ATCC 17029. We could not achieve an explanation for these two outliers.

Comparative and phylogenetic analyses of mammalian genomes facilitate our understanding of the underlying basis of disease-related and healthy phenotypes (Murphy et al., 2001; Li et al., 2013). Due to very large and complex genome structures, the alignment of whole chromosomal regions from more than a few species is not yet possible. Digital signal processing techniques, such as Fourier transform, provide novel approaches for comparative analysis of complex mammalian genomes. One possible solution is that we may divide the entire genomes into small segments, on which Fourier transform analysis can be applied. The distributions of Fourier power spectra on these segments can be used for genome analysis. In addition, mammalian genomes contain complex structural elements including tandem repeats, reverse repeats, transposable elements (TEs), exon, introns, and long non-coding RNA (Lnc RNA) genes (Jurka et al., 2007). These special elements often display periodic features, for example, exon sequences have distinct 3-periodicity and tandem and reverse repeats are periodic sequences. Because the Fourier transform can capture periodic signals in genomes, we will investigate in detail the applications of Fourier based algorithm on comparative analysis of large genomes, especially to address the impacts of these periodic genomic elements on phylogeny of genomes.

Determining the functional implications of gene or proteins sequences is one of the key tasks of the post-genome era. From similarity comparison and hierarchical clustering, we can infer functions and classify a new sequence or genome. This requires accurate and efficient similarity measure for DNA sequences. Most alignment-free methods such as the *k*-mer method and feature based methods may lose information after extracting sequence or feature information. The Fourier power spectrum makes a reversible comprehensive map and characterization of a DNA sequence and thus retains all the sequence information for comparison.

In this improved model on the similarity analysis of DNA sequences, the 2D mapping algorithm reduces the nucleotide composition bias in phylogenetic analysis. The even scaling method overcomes the limitation which requires the shortest length of a DNA more than one-half of the maximum length of the DNA compared and thus the DFT spectra of the short length DNA sequences can be evenly scaled to any length. The implementation of the new even scaling and reduction of dimensions representation in DNA sequence significantly reduces the computational complexity demands. Therefore, the improved DFT method is fast, accurate and low-complexity for comparing DNA sequences. Because of high efficiency and accuracy of the proposed DFT method, it can be used in whole genome phylogenetic analysis, which circumvents the ambiguity of choosing the genes for reconstruction and also avoids the necessity of aligning sequences of essentially different length and gene content. In addition, unlike other alignment-free methods, our DFT method does not need user defined parameters such as mer length in *k*-mer method, this feature makes DFT method be effective for

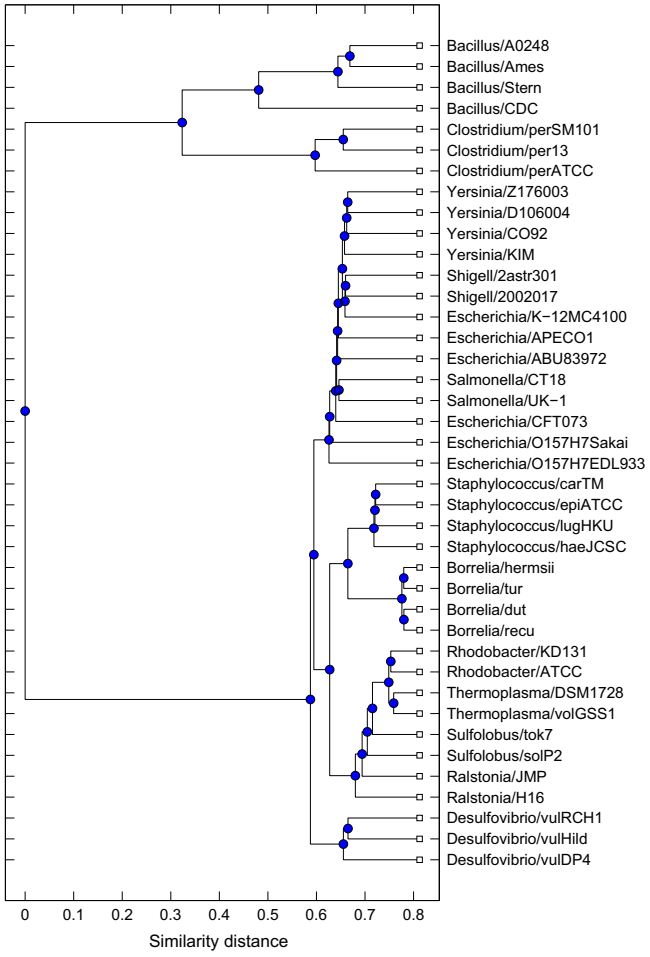

Fig. 9. Phylogenetic tree of bacterial genomes by the DFT distances of DNA sequences with the 2D numerical mapping. The size range of the bacterial genomes is 910k–5.5M bp.

analyzing new genomes when no taxon information is available. The numerical experiments on different DNA sequences and genomes with different sizes demonstrate the efficiency of the proposed method. Thus the improvement of the DFT method over our previous one is substantial.

High throughput sequencing is now fast and cost-effective, thus whole-genome sequencing will be available as a routine tool for clinical microbiology in the near future (Loman et al., 2012). Phylogenetic analysis of whole bacterial genomes is a critical tool for analyzing bacterial pathogenicity. When our algorithm for whole bacterial genome analysis is applied directly to clinical microbiology samples, it can significantly increase diagnostic accuracy and reduce processing time, thereby improve disease control and treatment. Moreover, from the public health perspective, our algorithm can be used in identifying the emergence of new threats and monitoring the spread of bacterial pathogens, and thus has great potentials in epidemiological investigations.

4. Conclusion

In this work, we establish an improved similarity measure based on Fourier transformation and even scaling for different length sequence data. The method has been assessed for accuracy by computer simulations and construction of phylogenetic trees of different virus genomes and genes. In this method, we first convert symbolic DNA sequences to 2D numerical sequences using nucleotide composition of the sequences, then we apply DFT on the 2D numerical sequences, and even scale the corresponding Fourier spectra to the longest sequences. The Euclidean distance was used to calculate the similarity of the scaled power spectrum in the same dimensional space. We created different DNA sequence mutants and assessed the accuracy of the new DFT metric on the mutants. The similarity metrics have been evaluated by constructing phylogenetic trees using different types of DNA sequences. The results show that the DFT based alignment-free DFT similarity measure provides highly accurate and computationally efficient identification of differences caused by a variety of mutants (point mutations, insertions/deletions and transposition) in DNA sequences. The DFT similarity measure method is a new effective tool in DNA sequence analysis at both gene and whole genome level.

Acknowledgment

This research is supported by the USA Natural Science Foundation (DMS-1120824 to S.S.-T. Yau) and National Natural Sciences Foundation of China (31271408 to S.S.-T. Yau), Tsinghua University startup fund and TsinghuaUniversity Independent Research Project grant. We are grateful to Prof. Jiasong Wang for helpful discussion, and we also thank Tung Hoang, Hui Zheng and Xuemeng E. Yin for proof reading the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this paper can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2015.06.033.

References

Agrawal, R., Faloutsos, C., Swami, A., 1993. Efficient similarity search in sequence databases. Springer, Berlin, Heidelberg.
Batista, G.E., Wang, X., Keogh, E.J., 2011. A complexity-invariant distance measure for time series. In: SDM, vol. 11, pp. 699–710.
Blaisdell, B.E., 1986. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc. Natl. Acad. Sci. 83 (14), 5155–5159.
Blaisdell, B.E., 1989. Effectiveness of measures requiring and not requiring prior sequence alignment for estimating the dissimilarity of natural sequences. J. Mol. Evol. 29 (6), 526–537.
Boore, J.L., 1999. Animal mitochondrial genomes. Nucleic Acids Res. 27 (8), 1767–1780.
Brisson-Noel, A., Arthur, M., Courvalin, P., 1988. Evidence for natural gene transfer from gram-positive cocci to *Escherichia coli*. J. Bacteriol. 170 (4), 1739–1745.
Chan, C.X., Ragan, M.A., 2013. Next-generation phylogenomics. Biol. Direct 8 (3).
Dai, Q., Li, Y., Liu, X., Yao, Y., Cao, Y., He, P., 2013. Comparison study on statistical features of predicted secondary structures for protein structural class prediction: from content to position. BMC Bioinform. 14 (1), 152.
Edgar, R.C., Batzoglou, S., 2006. Multiple sequence alignment. Curr. Opin. Struct. Biol. 16 (3), 368–373.
Eisen, J.A., 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res. 8 (3), 163–167.
Faloutsos, C., Ranganathan, M., Manolopoulos, Y., 1994. Fast subsequence matching in time-series databases, ACM New York. 23 (2), 419–429.
Foster, P.G., Hickey, D.A., 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J. Mol. Evol. 48 (3), 284–290.
Hildebrand, F., Meyer, A., Eyre-Walker, A., 2010. Evidence of selection upon genomic gc-content in bacteria. PLoS Genet. 6 (9), e1001107.
Jeffrey, H.J., 1990. Chaos game representation of gene structure. Nucleic Acids Res. 18 (8), 2163–2170.
Jun, S.-R., Sims, G.E., Wu, G.A., Kim, S.-H., 2010. 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.
Jurka, J., Kapitonov, V.V., Kohany, O., Jurka, M.V., 2007. Repetitive sequences in complex genomes: structure and evolution. Annu. Rev. Genomics Hum. Genet. 8, 241–259.
Kantorovitz, M.R., Robinson, G.E., Sinha, S., 2007. A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics 23 (13), i249–i255.

Kemena, C., Notredame, C., 2009. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. Bioinformatics 25 (19), 2455–2465.
Li, G., Davis, B.W., Raudsepp, T., Wilkerson, A.J.P., Mason, V.C., Ferguson-Smith, M., O'Brien, P.C., Waters, P.D., Murphy, W.J., 2013. Comparative analysis of mammalian Y chromosomes illuminates ancestral structure and lineage-specific evolution. Genome Res. 23 (9), 1486–1495.
Loman, N.J., Constantinidou, C., Chan, J.Z., Halachev, M., Sergeant, M., Penn, C.W., Robinson, E.R, Pallen, M.J, 2012. High-throughput bacterial genome sequencing: an embarrassment of choice a world of opportunity. Nat. Rev. Microbiol. 10 (9), 599–606.
Marhon, S.A., Kremer, S.C., 2011. Gene prediction based on DNA spectral analysis: a literature review. J. Comput. Biol. 18 (4), 639–676.
Marsella, L., Sirocco, F., Trovato, A., Seno, F., Tosatto, S.C., 2009. Repetita: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform. Bioinformatics 25 (12), i289–i295.
Mazodier, P., Davies, J., 1991. Gene transfer between distantly related bacteria. Annu. Rev. Genet. 25 (1), 147–171.
Mrázek, J., 2009. Phylogenetic signals in DNA composition: Limitations and prospects. Mol. Biol. Evol. 26 (5), 1163–1169.
Murphy, W.J., Stanyon, R., O'Brien, S.J., 2001. Evolution of mammalian genome organization inferred from comparative gene mapping. Genome Biol. 2 (6), 1–8.
Patil, K.R., McHardy, A.C., 2013. Alignment-free genome tree inference by learning group-specific distance metrics. Genome Biol. Evol. 5 (8), 1470–1484.
Qi, J., Wang, B., Hao, B.-I., 2004. Whole proteome prokaryote phylogeny without sequence alignment: a k-string composition approach. J. Mol. Evol. 58 (1), 1–11.
Qi, X., Wu, Q., Zhang, Y., Fuller, E., Zhang, C.-Q., 2011. A novel model for DNA sequence similarity analysis based on graph theory. Evol. Bioinform. Online 7, 149.
Rafiei, D., Mendelzon, A., 1998. Efficient retrieval of similar time sequences using dft. arXiv:⟨http://arXiv.org/abs/cs/9809033⟩.
Sharma, D., Issac, B., Raghava, G., Ramaswamy, R., 2004. Spectral repeat finder (srf): identification of repetitive sequences using fourier transformation. Bioinformatics 20 (9), 1405–1412.
Sims, G.E., Jun, S.-R., Wu, G.A., Kim, S.-H., 2009. Alignment-free genome comparison with feature frequency profiles (ffp) and optimal resolutions. Proc. Natl. Acad. Sci. 106 (8), 2677–2682.
Song, K., Ren, J., Reinert, G., Deng, M., Waterman, M.S., Sun, F., 2013. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. Brief. Bioinform., 343–353.
Sourdis, J., Krimbas, C., 1987. Accuracy of phylogenetic trees estimated from DNA sequence data. Mol. Biol. Evol. 4 (2), 159–166.
Tobe, S.S., Kitchener, A.C., Linacre, A.M., 2010. Reconstructing mammalian phylogenies: a detailed comparison of the cytochrome B and cytochrome oxidase subunit I mitochondrial genes. PLoS One 5 (11), e14156.
Vijaykrishna, D., Poon, L., Zhu, H., Ma, S., Li, O., Cheung, C., Smith, G., Peiris, J., Guan, Y., 2010. Reassortment of pandemic H1N1/2009 influenza A virus in swine. Science 328 (5985), 1529.
Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison: review. Bioinformatics 19 (4), 513–523.
Voss, R., 1992. Evolution of long-range fractal correlation and 1/f noise in DNA base sequences. Phys. Rev. Lett. 68, 3805–3808.
Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., Keogh, E., 2013. Experimental comparison of representation methods and distance measures for time series data. Data Min. Knowl. Discov. 26 (2), 275–309.
Wang, Y., Hill, K., Singh, S., Kari, L., 2005. The spectrum of genomic signatures: from dinucleotides to chaos game representation. Gene 346, 173–185.
Warnow, T., 2013. Large-scale multiple sequence alignment and phylogeny estimation, Models and Algorithms for Genome Evolution. Springer, Berlin, Heidelberg, pp. 85–146.
Wheeler, W.C., 1993. The triangle inequality and character analysis. Mol. Biol. Evol. 10, 707.
Wu, Y.-L., Agrawal, D., El Abbadi, A., 2000. A comparison of DFT and DWT based similarity search in time-series databases. In: Proceedings of the Ninth International Conference on Information and Knowledge Management, ACM, New York, pp. 488–495.
Xiong, C., Zhang, Z., Jiang, Q., Chen, Y. 2013. Evolutionary characteristics of A/Hangzhou/1/2013 and source of avian influenza virus H7N9 subtype in China. *Clinical infectious diseases*, cit294.
Yin, C., Chen, Y., Yau, S.S.-T., 2014. A measure of DNA sequence similarity by Fourier transform with applications on hierarchical clustering. J. Theor. Biol. 359 (21), 18–28.
Yin, C., Yau, S.S.-T., 2005. A Fourier characteristic of coding sequences: Origins and a non-Fourier approximation. J. Comput. Biol. 12 (9), 1153–1165.
Yin, C., Yau, S.S.-T., 2007. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. J. Theor. Biol. 247 (4), 687–694.