Journal of Theoretical Biology 359 (2014) 18–28

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering

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

[a] *College of Information Systems and Technology, University of Phoenix, Chicago, IL 60601, USA*
[b] *Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China*

HIGHLIGHTS

- We propose a distance measure for DNA sequences by Discrete Fourier Transform.
- We propose a method in even scaling time series.
- We apply the Discrete Fourier Transform distance on clustering of DNA sequences.

ARTICLE INFO

Article history:
Received 3 January 2014
Received in revised form
22 April 2014
Accepted 29 May 2014
Available online 6 June 2014

Keywords:
Genome
Similarity distance
Phylogenetic trees
Even scaling

ABSTRACT

Multiple sequence alignment (MSA) is a prominent method for classification of DNA sequences, yet it is hampered with inherent limitations in computational complexity. Alignment-free methods have been developed over past decade for more efficient comparison and classification of DNA sequences than MSA. However, most alignment-free methods may lose structural and functional information of DNA sequences because they are based on feature extractions. Therefore, they may not fully reflect the actual differences among DNA sequences. Alignment-free methods with information conservation are needed for more accurate comparison and classification of DNA sequences. We propose a new alignment-free similarity measure of DNA sequences using the Discrete Fourier Transform (DFT). In this method, we map DNA sequences into four binary indicator sequences and apply DFT to the indicator sequences to transform them into frequency domain. The Euclidean distance of full DFT power spectra of the DNA sequences is used as similarity distance metric. To compare the DFT power spectra of DNA sequences with different lengths, we propose an even scaling method to extend shorter DFT power spectra to equal the longest length of the sequences compared. After the DFT power spectra are evenly scaled, the DNA sequences are compared in the same DFT frequency space dimensionality. We assess the accuracy of the similarity metric in hierarchical clustering using simulated DNA and virus sequences. The results demonstrate that the DFT based method is an effective and accurate measure of DNA sequence similarity.

1. Introduction

With the advent of next generation sequencing technologies, a large volume post-genomic DNA sequence data are available, it has become increasingly important to develop effective and accurate similarity measure for comparing DNA sequence data. The discovery of novel biological knowledge from the *ab initio* analysis of DNA sequence data relies upon sequence comparison, classification, and clustering techniques. Phylogenetic tree analysis provides insights into the hierarchical relationships among genes, genomes and organisms, and thus becomes a fundamental research approach in structure comparison 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 pairwise distance measure of the DNA sequences using either multiple sequence alignment (MSA) or alignment-free methods on DNA sequences. The second phase is to construct the phylogenetic tree from the distance matrix using UPGMA or neighbor-joining tree construction method.

The traditional algorithms for comparing biological sequences are based mostly on sequence alignment. MSA plays a fundamental role in sequence comparison and is typically used to cluster

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

http://dx.doi.org/10.1016/j.jtbi.2014.05.043
0022-5193/© 2014 Elsevier Ltd. All rights reserved.

DNA and protein sequences (Warnow, 2013), but it has high computational complexity and requires large processing memory for long DNA sequences (Edgar and Batzoglou, 2006; Kemena and Notredame, 2009). In addition, most MSA methods try to minimize the number of insertion or deletion gaps in DNA sequences; therefore, MSA may create misalignments if the sequences contain weak homologous regions or mutations that involve longer segments of genomic sequences.

To overcome these problems in MSA, considerable researches on alignment-free methods have been developed. Blaisdell (1986) first proposed an alignment-free method based on the frequency of k-mer words of DNA sequences. This method is now widely used as an alignment-free method for genome analysis (Vinga and Almeida, 2003; Sims et al., 2009; Jun et al., 2010; Comin et al., 2012). The k-mer words in a DNA sequence are all possible permutations of length k from four nucleotides A, T, C, G. For example, if $k=5$, there are $4^5=1024$ such possible 5-mer fragments. The k-mer words method constructs fixed-length feature vectors by counting the frequencies of occurrence of all k-mer words in DNA sequences. The pairwise distances of the k-mer frequency vectors of different DNA sequences are measured by different distance metrics such as the Euclidean distance (Blaisdell, 1989) and Mahalanobis distances (Wu et al., 1997), or by information content measures such as Kolmogorov complexity (Li et al., 2001) and Lempel–Ziv complexity (Otu and Sayood, 2003). Dai et al. (2011) studied numerical characteristics of word frequencies, proposed a novel similarity measure by both the word frequencies and overlapping structures of words, and added directly k-word distribution statistics to Markov model to improve the performance of the k-mer method (Dai et al., 2008). This method is successfully used in many applications in biological sequence analysis, however, those distances depend considerably on the parameter k, and how to choose the optimal k that is dependent on varied degrees of divergence in sequence data (Jun et al., 2010). It also needs to address the issue for high computation complexity due to large number of k-mer string matching and high dimension of resulted frequency vector for large k-mer sizes. More recent developments in sequence comparison employ statistical and graphic properties of DNA and protein sequences (Dai et al., 2013; Qi et al., 2010; Yu et al., 2010; Deng et al., 2011). Although current alignment-free methods may solve the problems that MSA brings up, these alignment-free methods often require high computation time and memory space when words size k is large. More importantly, these methods also lose information within DNA sequences and have limited accuracy in clustering sequence data. It is therefore of an advantage to derive a similarity distance directly from the full information contents of DNA sequences.

Discrete Fourier Transform (DFT), a broadly used digital processing approach, may reveal hidden periodicities after transforming data from time domain to frequency domain space. The DFT method has been extensively used to study periodicities and repetitive elements in DNA sequences, genomes and protein structures (Anastassiou, 2001; Marhon and Kremer, 2011; Sharma et al., 2004; Marsella et al., 2009). One of the main results from applications of DFT in DNA sequence studies was the 3-periodicity property in DNA sequences, which gives a prominent peak in the Fourier power spectrum of protein coding-regions at frequency $f=1/3$ (Anastassiou, 2001). The power spectrum at $f=1/3$, the characteristic of protein coding regions in DNA sequences, reflects the non-uniform distribution of nucleotides in the three codon positions in the sequences (Yin and Yau, 2005). The 3-periodicity property is used to recognize coding and non-coding regions in DNA sequences (Tiwari et al., 1997; Yin and Yau, 2007; Yin et al., 2006). Because the DFT spectrum of a DNA sequence reflects the distribution of nucleotides on different periodic positions, it not only reveals the periodicities but also offers different views of data in frequency domain space. Due to the fact that the power spectrum conserves energy levels of signal in frequency domain according to Parseval's Theorem (Agrawal et al., 1993), the DFT method has been employed in efficient similarity searching in time-series and sequence databases and thus has potential as an alignment-free method for hierarchical clustering genome sequences.

Time series clustering has become an important topic, particularly for similarity search amongst long time series such as those arising in bioinformatics. To measure the distance between two or more time series of different lengths in Euclidean space, Dynamic Time Warping (DTW) has been applied in time series comparisons to resolve the difficulty caused when clustering time series of varying lengths (Gupta et al., 2005). DTW is time computation expensive for comparing multiple sequences. To overcome the heterogeneous lengths problems in DNA sequences, the most prominent method is to use DTW (Kruskal, 1983). But DTW may generate false information. In addition, DTW is only applicable to time series and cannot be used to compare two sequences in frequency spaces, where global hidden periodicity and periodic domain structures can be revealed and compared. The Euclidean distance is the most common method for discerning similarity in time series clustering, and it requires that the time series being compared are of exactly the same length dimensionality. New methods to address the different length problems are critical to compare and match time series in the Euclidean distance space.

In this study, we employ DFT power spectrum as a similarity measure for DNA sequences. The similarity metric uses the full Fourier power spectra of DNA sequences due to information congruence between time and frequency domain of the DNA sequences. To solve the length heterogeneous problem in DFT spectra of different sequence lengths, we propose an even scaling approach to extend short sequence to long sequence before comparing the absolute distance of power spectra for DNA sequences of different lengths. We present comprehensive experiments demonstrating the applicability and effectiveness of the proposed method in the hierarchical clustering of a variety of DNA sequences and genomes.

2. Methods and algorithms

2.1. Numerical representations of DNA sequences by 4-D binary indicators

DNA molecules are composed of four linearly linked nucleotides: adenine (A), thymine (T), cytosine (C), and guanine (G). A DNA sequence can be represented as a permutation of four characters A, T, C, G at different lengths. The character strings of DNA molecules are mapped into one or more numerical sequences. One of the methods in literatures is to use binary indicator sequences (Voss, 1992). A DNA sequence, 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 the absence of four nucleotides, A, T, C, and G, at the n-th position, respectively. The indicator mapping of DNA sequences is defined as follows:

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

where $\alpha \in A, C, G, T, n=0, 1, 2, \ldots, N-1$. The four indicator sequences correspond to the appearance of the four nucleotides at each position of the DNA sequence. For example, the indicator sequence, $u_A(n)=0001010111\ldots$, indicates that the nucleotide A is present in the positions of 4, 6, 8, 9, and 10 of the DNA sequence.

2.2. Discrete Fourier Transform

Discrete Fourier Transform (DFT) is the transformation of N observation data (time domain) to N new values in frequency domain. DFT spectral analysis of DNA sequences may detect any latent or hidden periodical signal in the original sequences. It may discover approximate repeats that are difficult to detect by algorithms based on tandem repeat search. Let U_A, U_T, U_C, and U_G be the DFT of the binary sequences u_A, u_T, u_C, and u_G, respectively; the DFT of the numerical series u_x of length N is defined as

$$U_x(k) = \sum_{n=0}^{N-1} u_x(n) e^{-i(2\pi/N)kn} \tag{1}$$

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

$$PS(k) = \sum_{x \in \{A,T,C,G\}} |U_x(k)|^2, \quad k = 0, 1, 2, \ldots, N-1 \tag{2}$$

where $U[k]$ is the k-th DFT coefficient.

Fourier Transform gives a unique representation of the original underlying signal in frequency domain. The frequency domain vector $U_x(k)$ contains all the information about $u_x(n)$. Parseval's Theorem for Fourier Transforms implies equivalence in the energy levels of signal in time and frequency domain. This property is the main driving force behind the new distance metric using DFT.

The power spectrum at frequency $f=N/3$ of a DNA sequence depends on the variance of nucleotide distributions in the three codon positions (Yin and Yau, 2005). From the DFT definition of an indicator sequence, the power spectrum is large when the nucleotide has a significant tendency of appearing about every N/k positions. In particular, when $k = N/3$, namely α tends to appear at a certain codon position. This leads to a prominent peak at frequency $f=N/3$. In one aspect, the power spectrum of a DNA indicator sequence $PS(k)$ represents the nucleotide distributions in every k-th position of a DNA sequence (Fukushima et al., 2002). This factor may contribute significantly to discrimination of sequences when the distribution of nucleotides is non-even. Otherwise, when the power spectrum is plain, it still serves as a transform of indicator sequences which contains distribution information of nucleotides. Because the Fourier power spectrum contains nucleotide distribution information, we propose to employ the DFT power spectra as similarity metric for comparing 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 the beginning few frequencies or last few frequencies (Wu et al., 2000; Wang et al., 2013; Rafiei and Mendelzon, 1998), but this approach may lose information for sequence comparison. To overcome the above problem, we propose here the following even scaling method to transform DFT power spectrum 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 PS_N denote the original power spectrum of length N and PS_M denote the extended power spectrum of length M from even scaling of PS_N. The symbol $[\ldots]$ denotes rounding integer operation. The even scaling operation on the original power spectrum PS_N to PS_M is defined as follows:

$$PS_M(k) = \begin{cases} PS_{\mathrm{N}}(k), & \text{if } k = 0 \\ \dfrac{1}{p-q+1} \displaystyle\sum_{j=q}^{p} PS_N(j) & \\ \text{where } p = \left[\dfrac{kN}{M}\right],\ q = \left[\dfrac{(k-1)N}{M}\right] & \text{if } k = 1 : M-1 \text{ and } M < 2N \end{cases}$$

For even scaling, the new length M is determined according to the longest length of the DNA sequences in a data set; for example, when constructing a phylogenetic tree of genomes, each DFT power spectrum of genomes is evenly scaled to the longest length among the compared genomes. After even scaling, the DFT spectra of the DNA sequences being compared are fitted into a new M-dimensional genome space. The pairwise similarity distance between two DNA sequences is measured as the Euclidean distance in the scaled genome space. Though the proposed even scaling method is applied to power spectrum sequence, it may also be used in even scaling other time series.

2.4. Algorithm for computing pairwise DFT distances of DNA sequences

A metric $d(x, y)$ is a non-negative function on a set of pairs (x, y) of finite sequences over a fixed alphabet. A distance metric of DNA sequences can be used as a measure of the evolutionary change from the sequence x to y. The evolutionary changes are reversible, and the fewest number of evolutionary changes is from x to y directly. Therefore, a metric is reflective, symmetric and transitive (Waterman, 1976; Otu and Sayood, 2003). A metric space is a set X together with a metric d on it. For example, the set of real numbers $\mathbb{R}$ with the function $d(x, y) = |x-y|$ is a metric space. We have the following conditions that a true metric shall satisfy in the metric space:

(1) $d(x, y) \geq 0$ for all $x, y \in X$; moreover, $d(x, y) = 0$, if and only if $x=y$.
(2) $d(x, y) = d(y, x)$ for all $x, y \in X$.
(3) The triangle inequality is satisfied, i.e., $d(x, y) \leq d(x, z) + d(z, y)$ for all $x, y, z \in X$

The triangle inequality is a property 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 is nonmetric and is internally inconsistent (Wheeler, 1993), then we will use the metric definition to verify if the proposed DFT distance is a true metric for DNA sequences. 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; Yu et al., 2011). The *Euclidean metric* on $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is defined by the function d:

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

After even scaling the DFT spectrum, we measured the distances of DNA sequences using the Euclidean distance of the full DFT power spectra of the DNA sequences. The distance measure in Fourier frequency domain in this paper excludes the zeroth term in the power spectrum because it is just the sum of data, otherwise, it may affect the accuracy of the measure because the zeroth power spectrum value usually is much larger than the rest of the power spectrum. Since we embed all the DNA sequence information via their full power spectrum into the same Euclidean space, the metric we propose is true induced the Euclidean distance in this Euclidean space.

The algorithm for computing pairwise DFT distances of DNA sequences SEQ1, SEQ2, and SEQ3 is as follows.

Algorithm 1. Algorithm for computing pairwise distances of DNA sequences in Fourier frequency domain.

```
Data: DNA SEQ1(length N1), SEQ2(length N2), SEQ3(length M),
    with M > N1, M > N2
Result: Pairwise distance of SEQ1, SEQ2 and SEQ3
Steps
  1. Convert SEQ1, SEQ2, SEQ3 to binary indicator sequence BS1, BS2, BS3
  2. Compute Fourier power spectrum PS1, PS2, and PS3M from BS1, BS2, BS3
  3. Even scale PS1 as PS1M from length N1 to length M
  4. Even scale PS2 as PS2M from length N2 to length M
  5. Compute the Euclidean distance $d$(PS1M,PS2M), $d$(PS2M,PS3M), $d$(PS1M,PS3M) in an M-dimensional space
```

2.5. Construction of phylogenetic trees

For comparison purpose, we used the following similarity measures methods: (1) The proposed even scaled DFT similarity measure implemented in MATLAB R2011b. (2) Alignment-free *k*-mer words method: The pairwise distance of the *k*-mer frequency vectors of different DNA sequences was measured by the Euclidean distance. The *k*-mer words method used a mer size as 7 in all the tests. The *k*-mer words method used the implementation as MATLAB NACS toolbox v4.1 (Vinga and Almeida, 2003). (3) Pairwise sequence alignment 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 method was performed using MATLAB R2011b bioinformatics toolbox. (4) MSA using Clustal W multiple sequence alignments method and the Jukes–Cantor genetic distance: The Clustal W method was performed using MEGA 6.0 (Tamura et al., 2007; Thompson et al., 1994).

The phylogenetic trees were constructed from distance matrices using UPGMA tree construction method. The UPGMA tree works by building the phylogenic tree bottom up from its leaves for the given set of species. It is basically a clustering algorithm with each species forming a cluster first, then two smaller clusters of nodes are grouped together recursively until there is only one phylogenic tree which contains all the species.

The methods and algorithms in this paper were implemented in MATLAB language and are available from the URL: https://sites.google.com/site/jtb2014yin/.

3. Results and discussions

3.1. Comparison of sequence similarity from Fourier frequency domain

Protein-coding regions of a DNA sequence exhibit a 3-base periodicity due to the non-uniform distribution of nucleotides in the three codon positions. The 3-base periodicity is rarely observed in intron regions. This property has been used in identifying the locations of protein-coding genes in unannotated sequence (Ficket and Tung, 1992). Before assessing the effectiveness of the proposed similarity metric based on DFT power spectra in comparing DNA sequences, we evaluated the consistency of DFT power spectra by even scaling of DNA sequences which contain protein coding regions (exons) and intron regions. The even-scaling method was applied to the DFT power spectra of the sequences to evenly extend the spectra to longer sizes. Fig. 1 is the DFT power spectrum of an exon of 731 bp and its evenly scaled power spectrum to the new length of 1031 bp. Fig. 2 is the DFT power spectrum of the first intron of human being myeloid cell leukemia protein 1 (350 bp) and its even-scaled power spectrum to the new length of 600 bp. Figs. 1 and 2 show that evenly scaled DFT power spectra resemble the original spectra before scaling. The basic statistical values for the spectra after even scaling are similar. For example, the mean values for the intron power spectrum before and after even scaling are 262.5845 and 262.712, respectively; the standard deviations for the intron power spectrum before and after even scaling are 150.5198 and 127.4096, respectively. These results demonstrate strong signal consistency by the even scaling method in terms of 3-base periodicity signal in the exon sequence and random power spectrum signal in the intron sequence.

A common similarity measure between two DNA sequences is edit distance, which is defined as the minimum number of insertions, deletions or substitutions of nucleotides needed to transform one sequence into the other. The edit distance can be obtained by optimal alignment of DNA sequences. Because DFT is

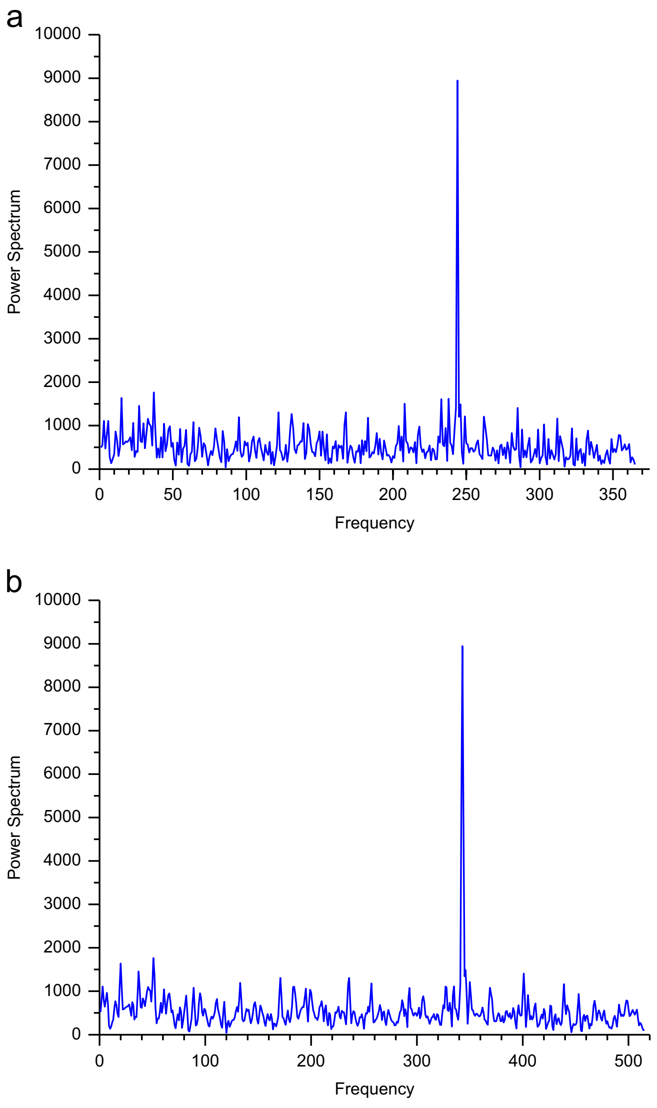

Fig. 1. DFT power spectrum of Bubo bubo voucher NHMO-BC120 cytochrome oxidase subunit 1 (COI) gene. The figures plot only the first half DFT spectrum of the gene. (a) Original DFT power spectrum and (b) even scaled DFT power spectrum. GenBank ID: GU571285.

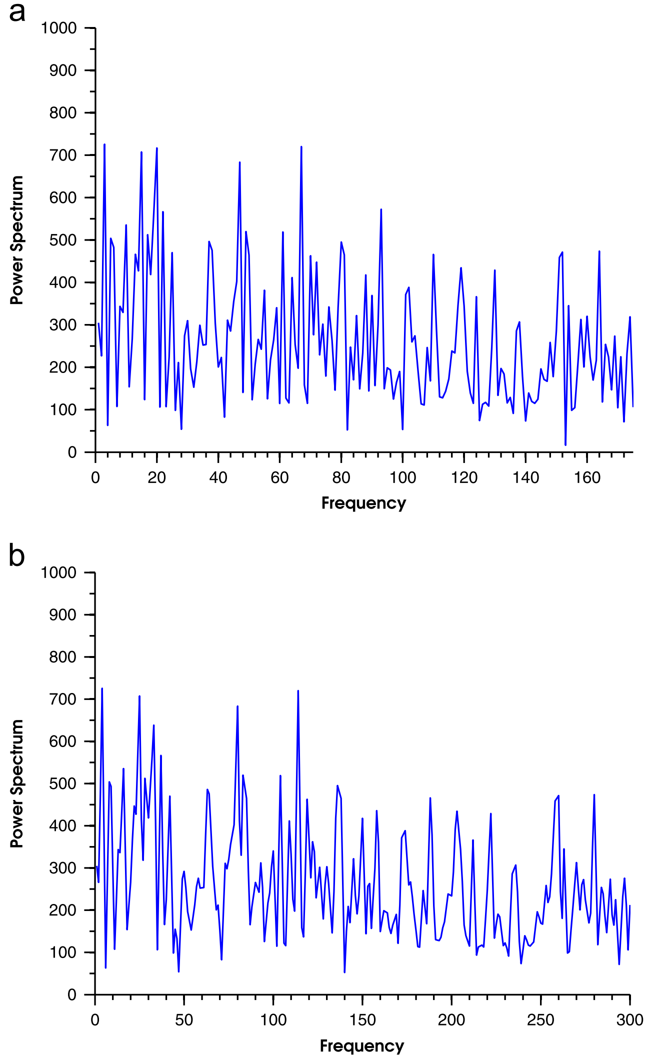

Fig. 2. DFT power spectrum of the first intron of myeloid cell leukemia protein 1 of *Homo sapiens* (*human*). The figures plot only the first half DFT spectrum of the gene. (a) Original DFT power spectrum and (b) even scaled DFT power spectrum. GenBank ID: AAG00896.

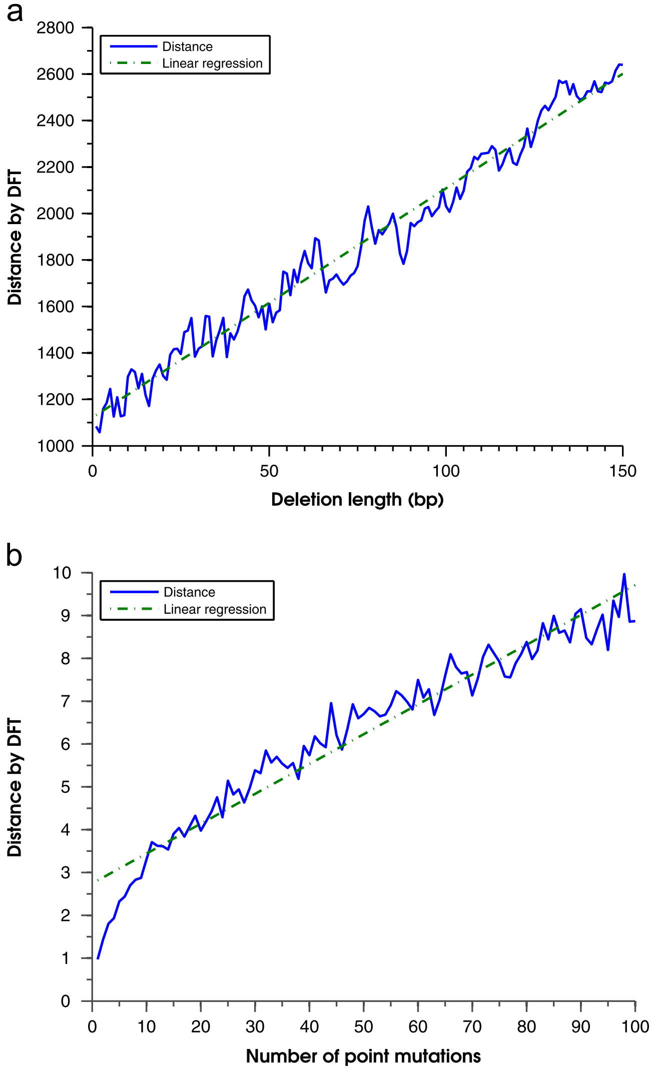

Fig. 3. (a) Correlation of the DFT distance and the lengths of deletion mutants of DNA sequences. (b) Correlation between DFT distance and the number of point mutations of DNA sequences.

an orthonormal time series transformation, which preserves lengths of vectors and angles between vectors, if we envision the input time series of length N as a vector in an N-dimensional space, applying DFT can be seen as a rotation of the space axes. These transformations do not affect the length of the original series according to Parseval's theorem, nor the Euclidean distance between any pair of the time series. Therefore, applying the Euclidean distance using all DFT coefficients gives the same classification performance as applying it using all original time features. Because we use full DFT spectrum in frequency domain, by Parseval's Theorem, the distance measure by DFT in frequency domain is expected to relate the edit distance in DNA sequences in time domain. Though the Euclidean distance of DFT power spectra in the same Euclidean space mathematically reflects the difference between two sequences, due to even scaling operation on the full power spectra DNA sequences, we use the following computational simulations to verify that the DFT distance of the scaled power spectrum truly reflects the similarity of the DNA sequences. To this end, we tested the correlation of the DFT distances and edit distances of deletions and substitutions on simulated DNA sequences.

We assessed the accuracy of the proposed similarity distance metric using a series of artificial deletion mutations of a DNA sequence and measured the correlation of the similarity distance and deletion sizes in these mutants. An intron sequence was partially deleted from 3′ end to generate different artificial mutants. The deletion size is from 1 bp to 100 bp. Then we measured the sequence distance between the mutants and the original sequence by the proposed DFT method. The result in Fig. 3 (a) is the correlation between the deletion lengths and the distances between the corresponding deletion mutants and original sequence. The result shows a sound linear relationship of DFT distances and the deletion mutations' lengths. This result shows 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 in DNA sequences. An intron sequence has introduced many 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 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 demonstrate an equivalency in DFT distance and edit distance for DNA sequence.

For a distance metric, the triangle inequality is a property of metric space. Distance that does not conform to this relation is nonmetric and is internally inconsistent. To verify if the distance metric satisfies the triangle property, we randomly selected 200 exons from the Exon–Intron Database (EID) (Shepelev and Fedorov, 2006) and measured pairwise distance of exons. For three randomly chosen exons as a test case, let d 1, d 2 and d 3 be the three distances measured in DFT frequency domain and d 3 be the largest distance in a test case. We compared the value of d 3 and d 1+d 2 to validate the triangle property. Fig. 4 shows that all the triangle property test cases satisfy the inequality, $d3 < d1 + d2$. The results demonstrate that the DFT-based distance is a valid distance measure.

3.2. Simulation of construction of phylogenetic trees on different DNA mutations

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 1 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 based method, alignment-free k-mer words method and pairwise sequence alignment method, as shown in Fig. 5(a), (b) and (c), respectively.

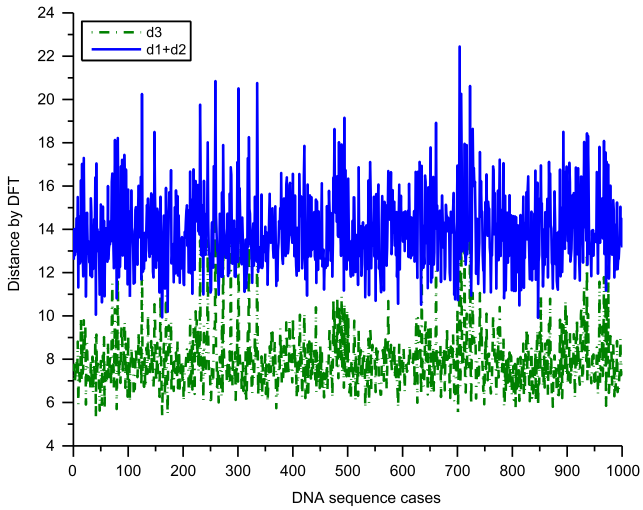

Fig. 4. Triangle property test of the DFT distances of DNA sequences.

For the different substitution mutations of the sequence A, Fig. 5(a)–(c) shows that the three methods can correctly classify and cluster them with correct tree topology. All the three methods create same tree topology corresponding to the numbers of substitution mutations in the DNA sequences. This indicates that the DFT similarity measure has the same capacity as the k-mer method and MSA method have to identify and measure the distances between substitutions. For deletion and insertion mutations of the sequence B, Fig. 5(a)–(c) shows the topological differences in DFT based measure and k-mer method and MSA method. Deletion and insertion are two serious mutations which are different from substitutions and most deletion and insertion may impact significant changes on phenotypes. Fig. 5(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 identify these deletion/insertion mutations from substitutions, mixing them in same branches (Fig. 5(b) and (c)). For transposition mutations, Fig. 5(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 detrimental mutations in hosts. Fig. 5(a) shows that 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. 5(b) and (c).

The possible cause of the problem is that the k-mer method may lose spatial positions information of nucleotides in DNA sequences. As an illustrative example, an insertion of nucleotide A into a short DNA sequence, *AACAAAACG*, at two different

Table 1
DNA sequence mutation description in simulation tests.

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

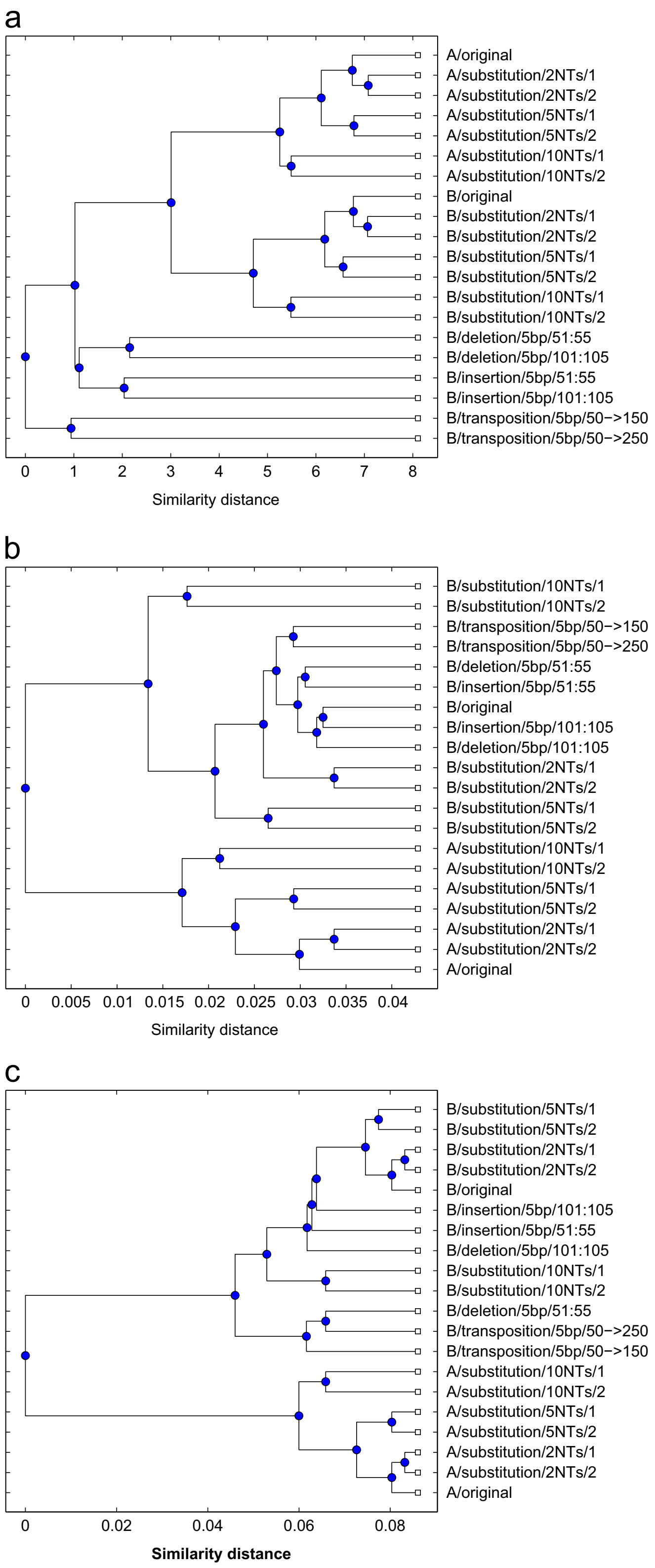

Fig. 5. Clustering analysis of different mutations by phylogenetic trees of simulated DNA sequences in Table 1. (a) The DFT distance, (b) the *k*-mer words, (c) pairwise sequence alignment.

positions, 1 or 6, the resulted insertion mutants are *AAACAAACG* and *AACAAAAAG*, but the two different mutants have the same *k*-mer frequency profile $[AAA, AAC, ACA, CAA, ACG, \ldots] = [2, 2, 1, 1, 1, 0, \ldots]$. MSA method only captures adjacent nucleotide sequence similarity and thus cannot recognize gene structure rearrangements. DFT based method reflects the nucleotide distribution on the positions in DNA sequences and can capture the fine characteristics of the sequences and thus recognize different types of mutations. Both *k*-mer method and MSA are mainly based on the orderings of nucleotides appearing in the sequence, but do not carry full position information of the sequences, thus similarity measures from *k*-mer and MSA are less reliable for sequence rearrangements. This result demonstrates that DFT similarity measure may have some special capacity to distinguish different mutations, whereas both *k*-mer and sequence alignment may miss these differences.

3.3. Construction of phylogenetic trees on individual genes

To test the utility of the proposed DFT distance measurement on individual genes, we used 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*, *Macaca mulatta*), one specie of new-world monkeys (*Saimiri scirueus*), two species of prosimians (*Lemur catta*, *Tarsisus syrichta*), and five hominoid species (Human, Chimpanzee, Gorilla, Orangutan and Hylobates) (Qi et al., 2010). In Fig. 6(a), the phylogenetic tree from DFT

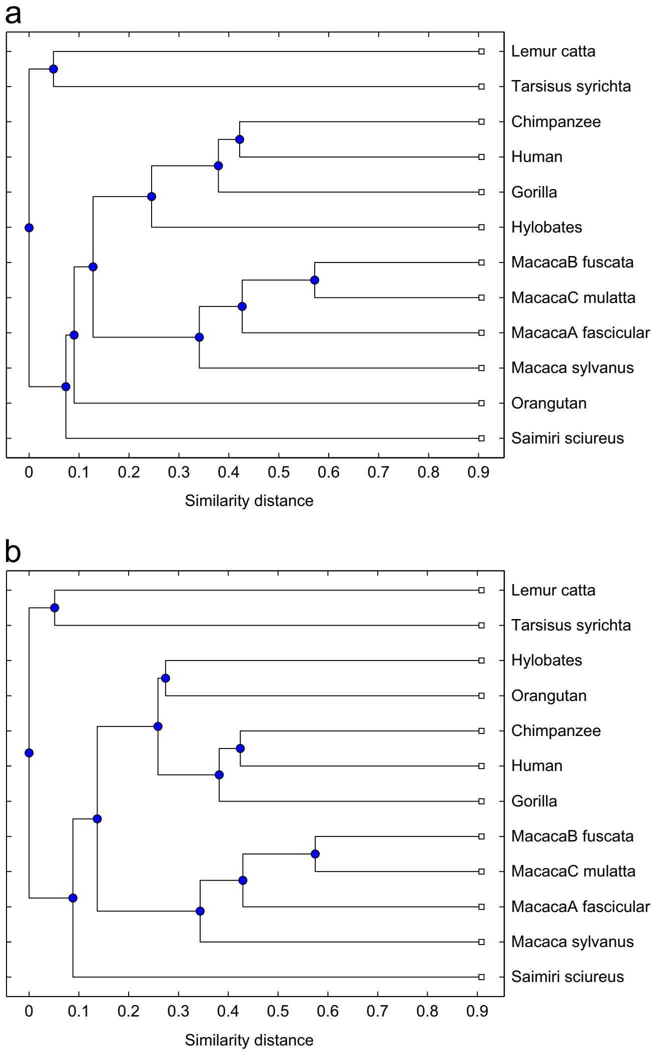

Fig. 6. Phylogenetic tree of 12 primate species by the DFT measure on NADH dehydrogenase subunit 4 gene. (a) Original 12 primates gene sequences, (b) Original 11 primates gene sequences and Orangutan with deletion mutation recovered by insertion C at position 558.

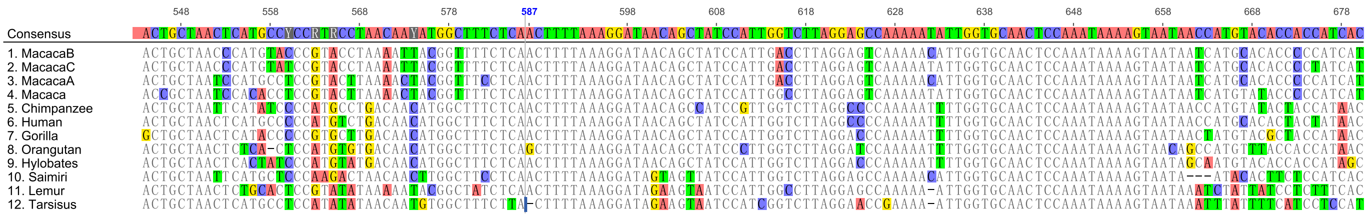

Fig. 7. Multiple sequence alignment of the NADH dehydrogenase subunit 4 from primates by Clustal W.

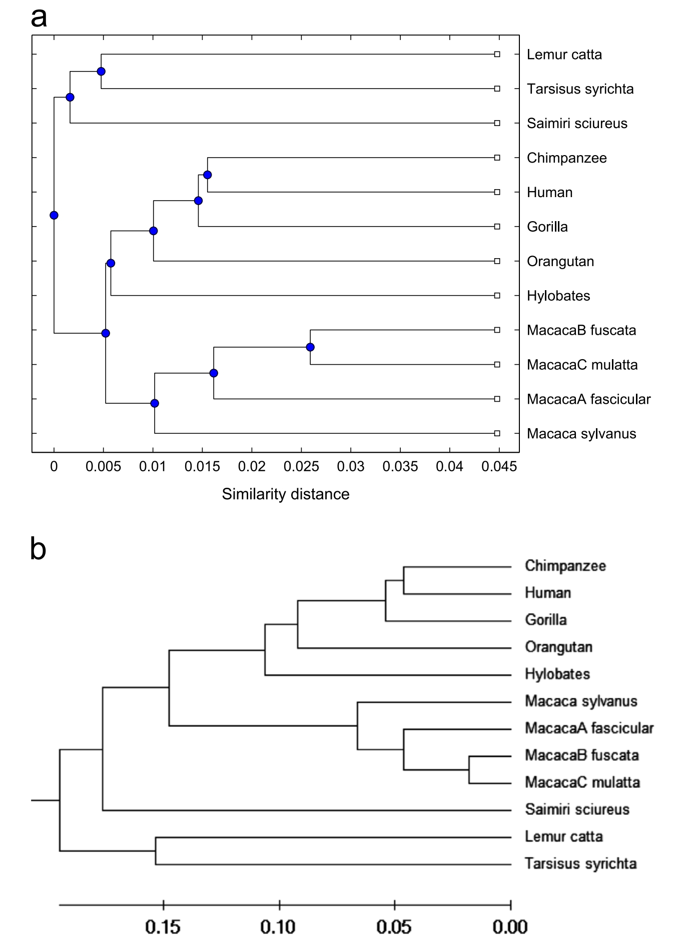

Fig. 8. Phylogenetic tree of 12 primate species from the DFT measure on NADH dehydrogenase subunit 4 gene. (a) *k*-mer method, and (b) MSA by Clustal W and Jukes–Cantor distance using MEGA 6.0.

method of these 12 species is generally consistent with the previous works with one exception that Orangutan is far from hominoid species. To investigate the reason of this exception, using sequence alignment by MEGA, we found that there is a base deletion mutation at position 558 in the NADH dehydrogenase subunit 4 of Orangutan compared with the segments from Human, Chimpanzee, and Gorilla (Fig. 7). If this deletion is recovered by inserting nucleotide C at position 558, the phylogenetic tree constructed from DFT method is the same as those from *k*-mer method and MSA method as shown in Figs. 6(b), and 8(a) and (b). The phylogenetic trees using the *k*-mer method and MSA method on the segments of Orangutan before and after deletion mutation recovery are the same. These results indicate that DFT based method can identify a single nucleotide deletion, but *k*-mer method and MSA method cannot recognize the difference of deletion mutation and its recovery. These results explain the differences of the phylogenetic trees by the DFT similarity measure, *k*-mer words method, and MSA method. This case study on real DNA sequence is consistent with previous simulation results, demonstrating that DFT based method can reveal the difference in phylogenetic tree caused by different mutations.

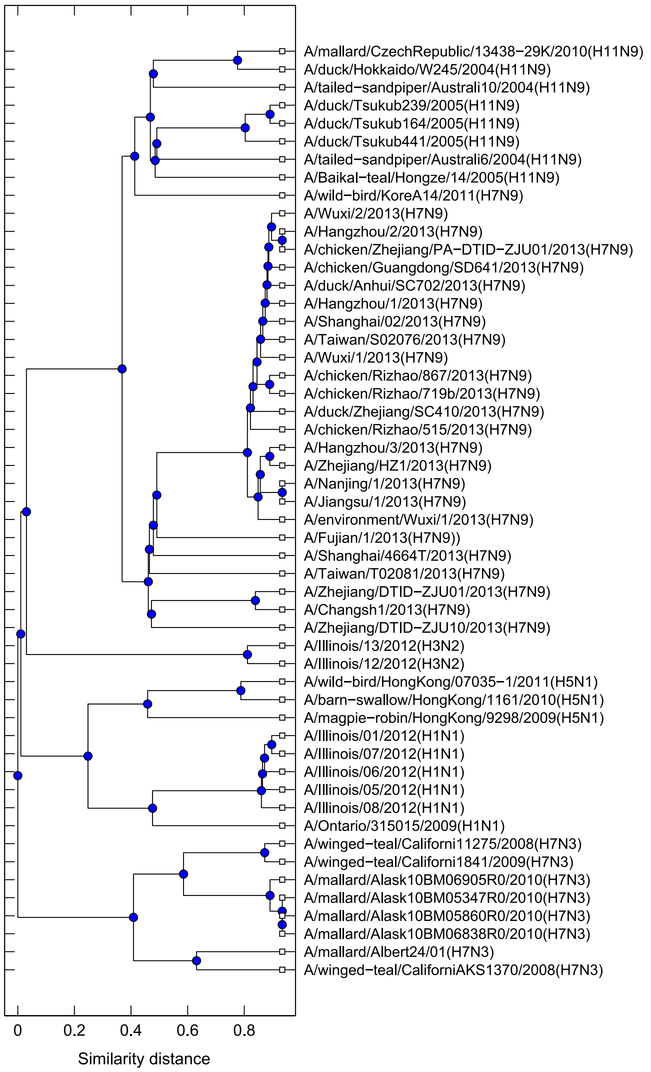

Fig. 9. Phylogenetic tree of influenza A viruses using the DFT distances.

We assessed the effectiveness of the DFT metric in measuring individual gene level. In the test, we used Influenza A virus neuraminidase (NA) gene because of its association with pandemic influenza and a wide range of natural hosts, including man, birds, and other animals. We constructed phylogenetic tree based on pairwise DFT distance of the segment 6 neuraminidase (NA) gene of different influenza A strains. Figs. 9 and 10 are the phylogenetic trees of influenza A virus constructed by the proposed DFT method and sequence alignment method with Jukes–Cantor distance, respectively. We used two different sequence alignment methods: one was from MATLAB 2011b bioinformatics toolbox and the other was from Clustal W in MEGA 6.0. Two sequence alignment

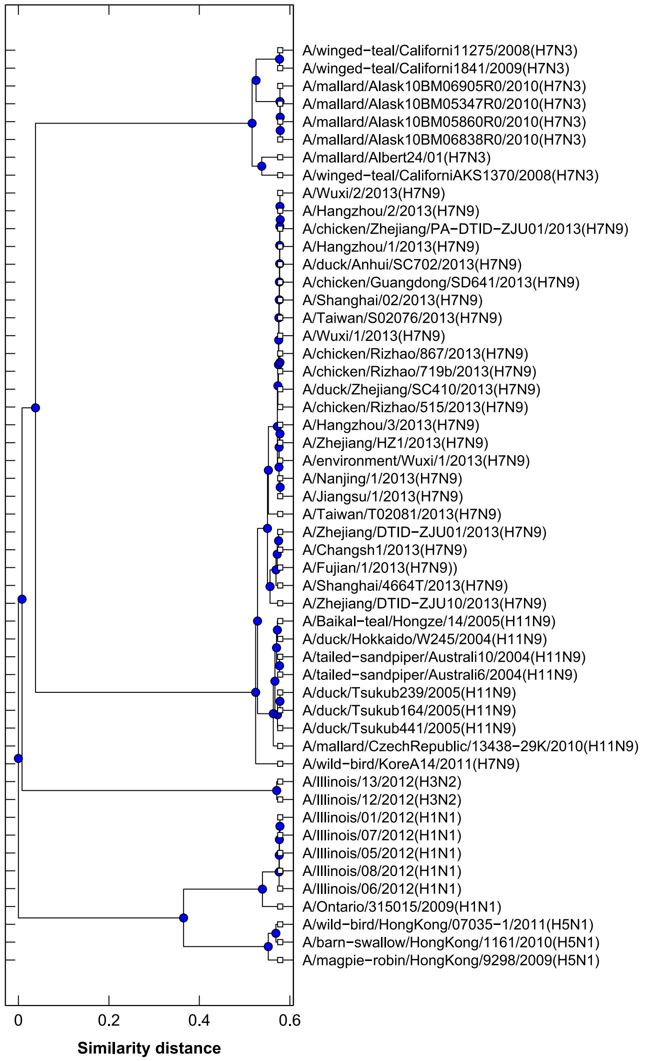

Fig. 10. Phylogenetic tree of influenza A viruses using MATLAB pairwise sequence alignment and Jukes–Cantor distances.

methods generated same results (Fig. 10 here and Fig. A.1 in supplementary material). Both trees show correct grouping of different virus subtypes H7N9, H11N9, H3N2, and H1N1. The tree from DFT distance shows clear branch difference than the tree from the Jukes–Cartor distance. The virus of highly homologous sequences such as A/Illinois H1N1 virus, 06/2012, 08/2012, and 01/2012, 07/2012 cannot be separated by sequence alignment measured by Jukes–Cartor method, but they are clearly separated with correct hierarchical relationship in the tree of DFT method. The other example in the figure is that the N7N9 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). Thus, the DFT tree can display clear levels of hierarchy and relationship among different viruses, but Jukes–Cantor cannot have clear spatial separation of similar species in the tree. These results demonstrate the superiority of the proposed DFT method on the existing sequence alignment methods due to the fact that the DFT distance is from calculation of all the sequence information and does not lose any sequence information after Fourier transform.

3.4. Construction of phylogenetic trees on whole genomes

We evaluated and applied the proposed DFT similarity measure on hierarchical clustering genomes, which contain different genes and non-coding regions. The test genomes were 80 different human rhinovirus (HRV) genomes. The DFT distances between any two HRV species were measured after even scaling each DFT spectrum to the longest genome size among all compared HRV genomes. The pairwise DFT distances were used to construct a similarity matrix in construction of the phylogenetic trees by UPGMA method (Sneath et al., 1973). To compare the effectiveness between DFT distance metric and sequence alignments in hierarchical clustering, we used the Jukes–Cantor sequence alignment model of DNA sequence evolution (Jukes and Cantor, 1969). The Jukes–Cantor method assumes that every site evolves independent of the others, so it suffices to analyze one site at a time. It also assumes that every base (i.e. the purines A and G and the pyrimidines C and T) has a constant probability per unit time of changing into each of the others bases. Fig. 11 is the phylogenetic tree of HRV genomes constructed by DFT distances. The GenBank access IDs for the virus are provided as supplementary material (Palmenberg et al., 2009). The result in Fig. 11 shows correct

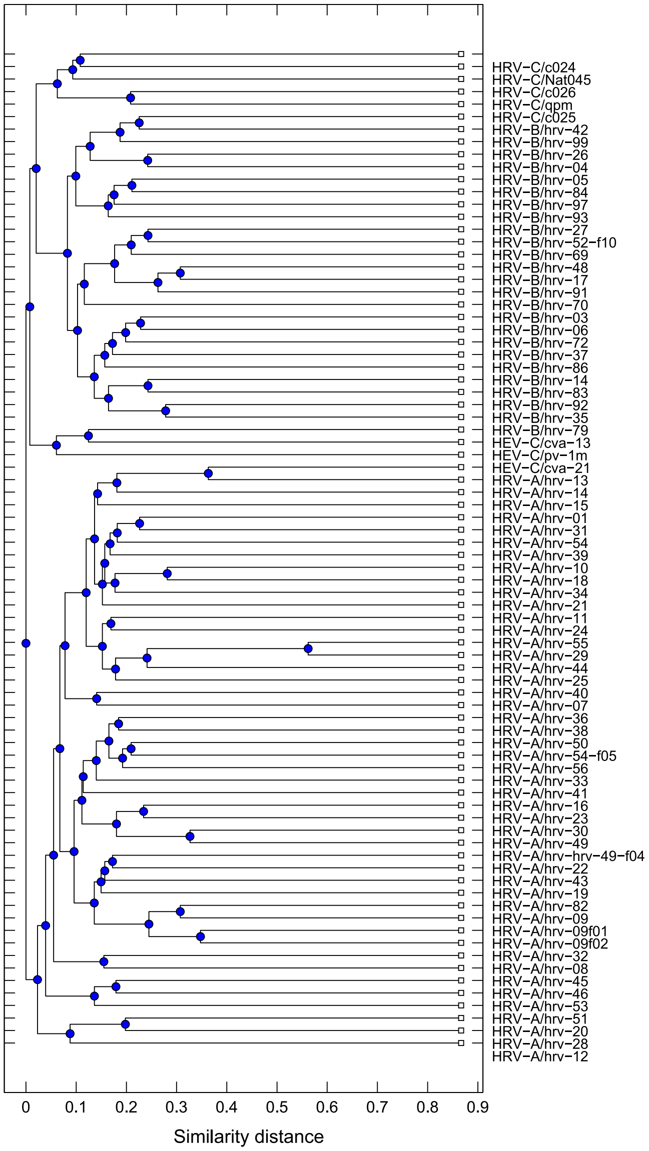

Fig. 11. Phylogenetic tree of HRV genomes by the DFT distances.

Table 2
Performance comparison by benchmark test on 80 HRV genomes.

Method	Time (min)	Parameters
DFT	19.50	Max length: 7458 bp
k-mer	74.13	NASC MATLAB toolbox, $k=7$
MSA	897.0	Clustal W

grouping of different virus types HRV A, B and C, and HEVC. The tree from the DFT method is consistent with *k*-mer and MSA methods (data from *k*-mer and MSA shown in supplementary materials) (Palmenberg et al., 2009). The results demonstrate that the DFT distance can be used successfully in comparing and classifying both individual genes and whole genomes.

We compared the performance of three methods, DFT distance measure, *k*-mer distance measure and MSA, on measuring the pairwise distances or align the sequences among the 80 HRV genomes. The performance was tested on the same hardware configurations. The benchmark tests in Table 2 show that the DFT method reduces 73.7% processing time of the *k*-mer method and achieves accurate results as in Fig. 11. Matching *k*-mer from large DNA sequence takes significant processing time and memory resource. This result is consistent with previous study (Melsted and Pritchard, 2011). The table also shows that MSA needs 13 h to align the same set of genomes. The time spent in MSA is much longer than DFT and *k*-mer methods. Though DFT achieves better performance compared with *k*-mer and MSA, it still has a relatively high computational complexity for very long DNA sequence such as large whole genomes. Computing DFT for large values of N is very intensive because we have N^2 complex multiplications for direct DFT. Even using the fast Fourier Transform (FFT) method, we still need $N \log N$ multiplications. Future study will be on DFT measures from non-overlapped segments of long DNA genome sequences. We will investigate to reduce computational complexity while minimizing difference between DFT distance and edit distance.

Another limitation of the DFT based method in DNA comparison is that if the shortest length of a DNA is less than one half of the maximum length of the DNA compared, the DFT spectrum of the shortest length DNA cannot be evenly scaled to the maximum length. We will address this limitation in future study.

One of the key tasks of the post-genome era is to determine the functional implications of gene or proteins sequences. From similarity comparison and hierarchical clustering, we may be able to infer functions and classify a new sequence or a 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 retain all the sequence information for comparison. The proposed DFT distance metric leads to reliable results in hierarchical clustering of DNA sequences and shows a better identification of different mutations in hierarchical tree and improves speed over the *k*-mer method and MSA.

4. Conclusion

In this work, we establish a new and robust distance measure method based on Fourier transformation and propose an even scaling method to compare different length data. The method has been assessed for accuracy by computer simulations and construction of phylogenetic trees of different virus genomes and genes. In the method, we first performed DFT on DNA sequences after converting symbolic sequences to four binary indicator sequences, then DFT spectra of different lengths were evenly scaled to the same length of the longest sequences. The Euclidean distance was used to calculate the similarity of the scaled power spectrum. 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 method provides highly accurate and computationally efficient identification of differences caused by a variety of mutants (point mutations, insertions/deletions and transposition) in DNA sequences. This study opens an avenue for future research into efficient DNA comparison algorithms for large genomes and short reads in next generation sequencing.

Acknowledgments

We thank Xuemeng E. Yin at William Fremd High School, Dr. Chenglong Yu and Tung T. Hoang at University of Illinois at Chicago for proof reading paper. We gratefully acknowledge Ms. Xiujuan Yu at Department of Mathematical Sciences, Tsinghua University, China, for her support on the project. We appreciate the two reviewers for their valuable suggestions, which greatly improved the quality of this paper. This research is supported by the USA Natural Science Foundation (DMS-1120824 to S. S.-T. Yau), National Natural Sciences Foundation of China (31271408 to S.S.-T. Yau), Tsinghua University start up fund (to S. S.-T. Yau) and Tsinghua University independent research project grant (to S. S.-T. Yau).

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.2014.05.043.

References

Agrawal, R., Faloutsos, C., Swami, A., 1993. Efficient similarity search in sequence databases. Springer, Berlin Heidelberg, pp. 69–84.

Anastassiou, D., 2001. Genomic signal processing. IEEE Signal Process. Mag. 18, 8–20.

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.

Comin, M., Verzotto, D., et al., 2012. Alignment-free phylogeny of whole genomes using underlying subwords. Algorithms Mol. Biol. 7 (1).

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.

Dai, Q., Liu, X., Yao, Y., Zhao, F., 2011. Numerical characteristics of word frequencies and their application to dissimilarity measure for sequence comparison. J. Theor. Biol. 276 (1), 174–180.

Dai, Q., Yang, Y., Wang, T., 2008. Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison. Bioinformatics 24 (20), 2296–2302.

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 (3), e17293.

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.

Ficket, J.W., Tung, C.S., 1992. Assessment of protein coding measure. Nucleic Acids Res. 20, 6441–6450.

Fukushima, A., Ikemura, T., Oshima, T., Mori, H., Kanaya, S., 2002. Detection of periodicity in eukaryotic genomes on the basis of power spectrum analysis. Genome Informatics Ser., 21–29.

Gupta, K., Thomas, D., Vidya, S., Venkatesh, K., Ramakumar, S., 2005. Detailed protein sequence alignment based on spectral similarity score (SSS). BMC Bioinform. 11, 112–122.

Jukes, T., Cantor, C., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.) Mammalian Protein Metabolism, pp. 21–132.

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.

Kemena, C., Notredame, C., 2009. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. Bioinformatics 25 (19), 2455–2465.

Kruskal, J., 1983. An overview of sequence comparison: time warps, string edits, and macromolecules. SIAM Rev. 25, 201–237.

Li, M., Badger, J.H., Chen, X., Kwong, S., Kearney, P., Zhang, H., 2001. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics 17 (2), 149–154.

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.

Melsted, P., Pritchard, J.K., 2011. Efficient counting of k-mers in dna sequences using a bloom filter. BMC Bioinform. 12 (1), 333.

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

Palmenberg, A.C., Spiro, D., Kuzmickas, R., Wang, S., Djikeng, A., Rathe, J.A., Fraser-Liggett, C.M., Liggett, S.B., 2009. Sequencing and analyses of all known human rhinovirus genomes reveal structure and evolution. Science 324 (5923), 55–59.

Qi, X., Wu, Q., Zhang, Y., Fuller, E., Zhang, C.-Q., 2010. A novel model for dna sequence similarity analysis based on graph theory. Evolut. Bioinform. Online 7, 149–158.

Rafiei, D., Mendelzon, A., 1998. Efficient Retrieval of Similar Time Sequences Using DFT. arXiv preprint 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.

Shepelev, V., Fedorov, A., 2006. Advances in the exon–intron database. Data Min. Knowl. Discov. 7, 178–185.

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.

Sneath, P.H., Sokal, R.R., et al., 1973. Numerical Taxonomy: The Principles and Practice of Numerical Classification.

Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24 (8), 1596–1599.

Thompson, J.D., Higgins, D.G., 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. Nucleic Acids Res. 22 (22), 4673–4680.

Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S., Ramaswamy, R., 1997. Prediction of probable genes by fourier analysis of genomic sequences. Bioinformatics 13 (3), 263–270.

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.

Warnow, T., 2013. Large-scale multiple sequence alignment and phylogeny estimation, Models and Algorithms for Genome Evolution. Springer, London, pp. 85–146.

Waterman, M., 1976. Some biological sequence metrics. Adv. Math. 20, 367–387.

Wheeler, W.C., 1993. The triangle inequality and character analysis. Mol. Biol. Evol. 10, 707.

Wu, T.-J., Burke, J.P., Davison, D.B., 1997. A measure of dna sequence dissimilarity based on mahalanobis distance between frequencies of words. Biometrics 53 (4), 1431–1439.

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 57 (4), 622–624.

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.

Yin, C., Yau, S.-T., 2005. A fourier characteristic of coding sequences: origins and a non-fourier approximation. J. Comput. Biol. 12 (9), 1153–1165.

Yin, C., Yoo, D., Yau, S.T. (2006). Tracking the 3-Base Periodicity of Protein-Coding Regions by the Nonlinear Tracking-Differentiator. Decision and Control, 2006 45th IEEE Conference on. IEEE, New York, pp. 2094–2097.

Yu, C., Cheng, S.-Y., He, R.L., Yau, S.S.-T., 2011. Protein map: an alignment-free sequence comparison method based on various properties of amino acids. Gene 486 (1), 110–118.

Yu, C., Liang, Q., Yin, C., He, R.L., Yau, S.S.-T., 2010. A novel construction of genome space with biological geometry. DNA Res. 17 (3), 155–168.