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.