Journal of Theoretical Biology 372 (2015) 135–145

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

A new method to cluster DNA sequences using Fourier power spectrum

Tung Hoang [a], Changchuan Yin [a], Hui Zheng [a], Chenglong Yu [b,c], Rong Lucy He [d], Stephen S.-T. Yau [e,*]

[a] Department of Mathematics, Statistics and Computer Science, University of Ilinois at Chicago, Chicago, IL 60607, USA
[b] Mind and Brain Theme, South Australian Health and Medical Research Institute, North Terrace, Adelaide, SA 5000, Australia
[c] School of Medicine, Flinders University, Adelaide, SA 5001, Australia
[d] Department of Biological Sciences, Chicago State University, Chicago, IL, USA
[e] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China

HIGHLIGHTS

- We propose to use Fourier power spectrum to cluster genes and genomes.
- We construct mathematical moments from the power spectrum.
- We perform phylogenetic analysis of genes and genomes based on moments.

ARTICLE INFO

Article history:
Received 12 September 2014
Received in revised form
15 January 2015
Accepted 23 February 2015
Available online 5 March 2015

Keywords:
Phylogenetic trees
Genes
Moments

ABSTRACT

A novel clustering method is proposed to classify genes and genomes. For a given DNA sequence, a binary indicator sequence of each nucleotide is constructed, and Discrete Fourier Transform is applied on these four sequences to attain respective power spectra. Mathematical moments are built from these spectra, and multidimensional vectors of real numbers are constructed from these moments. Cluster analysis is then performed in order to determine the evolutionary relationship between DNA sequences. The novelty of this method is that sequences with different lengths can be compared easily via the use of power spectra and moments. Experimental results on various datasets show that the proposed method provides an efficient tool to classify genes and genomes. It not only gives comparable results but also is remarkably faster than other multiple sequence alignment and alignment-free methods.

1. Introduction

In the last few decades, several methods to classify genes and proteins have been proposed. Most of these methods are alignment-based in which optimal alignments are obtained by using selected scoring systems. These methods provide accurate classification of biological sequences, and several algorithms have been developed and successfully applied (Katoh et al., 2002; Edgar, 2004; Larkin et al., 2007). Nevertheless, their major drawback is due to significantly high time and memory consumption which is not suitable when a quick clustering needs to be made, for example on a new deadly virus (Marra et al., 2003). Henceforth, an alignment-free technique is a trending method that often gives much faster classification on the same dataset (Vinga and Almeida, 2003; Yau et al., 2008; Yu et al., 2011, 2013). For example, the *k*-mer method is among the most popular alignment-free methods. In order to measure how different the two sequences are, the set of *k*-mers, or subsequences of length *k*, in the two biological sequences are collected and then the evolutionary distance between them is computed (Vinga and Almeida, 2003; Pandit and Sinha, 2010). The *k*-mer method gives comparable results to alignment-based methods while being computationally faster (Blaisdell, 1989).

Discrete Fourier Transform (DFT) is a powerful tool in signal and image processing. During recent years, DFT has been increasingly used in DNA research, such as gene prediction, protein coding region, genomic signature, hierarchical clustering, periodicity analysis (Tiwari et al., 1997; Anastassiou, 2000; Kotlar and Lavner, 2003; Vaidyanathan and Yoon, 2004; Afreixo et al., 2004, 2009; Tenreiro Machado et al., 2011). A DFT power spectrum of a DNA sequence reflects the nucleotide distribution and periodic

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

http://dx.doi.org/10.1016/j.jtbi.2015.02.026
0022-5193/

patterns of that sequence, and it has been applied to identify protein coding regions in genomic sequences (Fukushima et al., 2002; Yin and Yau, 2005, 2007). In this paper we provide a new alignment-free method to classify DNA sequences based on the DFT power spectrum. The method is tested and compared to other state-of-the-art methods on various datasets for speed and accuracy.

2. Materials and method

2.1. Mathematical background

In signal processing, sequences in time domain are commonly transformed into frequency domain to make some important features visible. Via that transformation, no information is lost but some hidden properties could be revealed (Oppenheim et al., 1989).

One of the most common transformations is Discrete Fourier Transform (Oppenheim et al., 1989). For a signal of length $N, f(n), n=0,\ldots,N-1$, the DFT of the signal at frequency k is

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

for $k=0,\ldots,N-1$. The DFT power spectrum of a signal at frequency k is defined as

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

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

The DFT is often used to find the frequency components of a signal buried in a noisy time domain. For example, let y be a signal containing a 60 Hz sinusoid of amplitude 0.8 and a 140 Hz sinusoid of amplitude 1. This signal can be corrupted by a zero-mean random noise:

$$y=0.8*\sin(2*\pi*60*t)+\sin(2*\pi*140*t)+\mathrm{random}$$

The frequencies can hardly be identified by looking at the original signal as in Fig. 1(a), but can be seen quite clearly when the signal is transformed to frequency domain by taking the DFT (Fig. 1(b)).

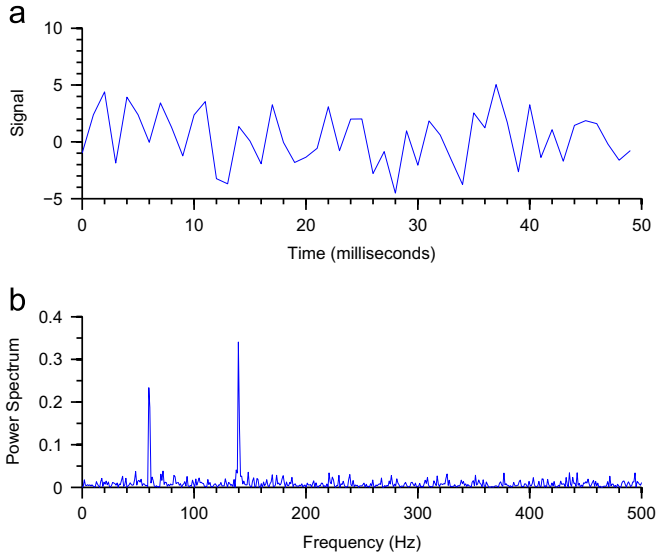

Fig. 1. Signal in time domain and frequency domain. (a) Signal Corrupted with Zero – Mean Random Noise. and (b) Single – Sided Power Spectrum

2.2. Moment vectors

For a DNA sequence composed of nucleotides adenine (A), cytosine (C), guanine (G), and thymine (T), one typical way to get numerical representation is to use binary indicator sequences. The values of these sequences are either 0 or 1 indicating the absence or presence of a specific nucleotide. Specifically, for a given DNA sequence of length N, we define u_A of the same length as follows:

$$u_A(n)=\begin{cases}1 & \text{if } A \text{ is present at location } n \text{ of the sequence}\\ 0 & \text{otherwise}\end{cases}$$

u_C, u_G, u_T are defined similarly.

For example, for the sequence *AGTCTTACGA*, the corresponding indicator sequence of nucleotide A is $u_A=1000001001$.

The DFT of u_A is U_A where

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

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

The DFT power spectrum of u_A is PS_A where $\mathrm{PS}_A(k)=|U_A(k)|^2, k=0,\ldots,N-1$. The corresponding power spectrum for nucleotides C, G, T is defined similarly. In general, for a gene sequence of length N, let N_A, N_C, N_G, N_T be the number of nucleotide A, C, G, T in that sequence, respectively.

It is difficult to compare numerical sequences with different lengths, so we cannot cluster genes and genomes based on their power spectra sequences. One common approach to get over this problem is to use mathematical moments, e.g. for nucleotide A defines jth moment $M_j^A=\alpha_j^A\sum_{k=0}^{N-1}(\mathrm{PS}_A(k))^j, j=1,2\ldots$, where α_j^A be scaling factors. We want higher moments to converge to zero, i.e. essential information is kept in the first few moments. Thus, the chosen normalization factors α_j^A must reflect the nature of the sequences. Let us examine the binary indicator sequence of one nucleotide, A, in more detail.

By Parseval's theorem (Oppenheim et al., 1989),

$$\sum_{n=0}^{N-1}|u_A(n)|^2=\frac{1}{N}\sum_{k=0}^{N-1}\mathrm{PS}_A(k)\quad(\text{since } \mathrm{PS}_A(k)=|U_A(k)|^2)$$

The left side is actually N_A, i.e. the number of 1 in the A binary sequence. Hence, $\sum_{k=0}^{N-1}\mathrm{PS}_A(k)=N_AN$. So it is reasonable for α_j^A to be a power of N_{AN}. As stated above, we want moments converge to zero gradually so that information loss is minimal, thus $\alpha_j^A=1/(N_AN)^{j-1}$ is the best choice (which will be verified later). Therefore

$$M_j^A=\frac{1}{N_A^{j-1}N^{j-1}}\sum_{k=0}^{N-1}(\mathrm{PS}_A(k))^j$$

With this normalization, $M_1^A=\sum_{k=0}^{N-1}\mathrm{PS}_A(k)=N_AN$. Our experimental results on various datasets proved that this is a good normalization. However, by re-examining the formula, we find that a slight modification can be made to get better outcomes. From Section 2.1, we know $\mathrm{PS}_A(0)=|F_A(0)|^2=|\sum_{n=0}^{N-1}u_A(n)|^2=N_A^2$. Thus $\mathrm{PS}_A(0)$

Table 1
Running time comparison.

Datasets	Our method	MAFFT	*k*-mer	ClustalW
Mammals	4 s	NA	18 min 15 s	3 h 15 min
Influenza A	0.6 s	22 s	12 s	1 min 55 s
HRV	5 s	17 min 40 s	47 min 28 s	8 h 10 min
Coronavirus	6 s	NA	69 min 12 s	11 h 40 min
Bacteria	9 min 41 s	NA	NA	NA