Analyzing Exon Structure with PCA and ICA of Short-Time Fourier Transform

Changha Hwang[1], Insuk Sohn[2]

Abstract

We use principal component analysis (PCA) to identify exons of a gene and further analyze their internal structures. The PCA is conducted on the short-time Fourier transform (STFT) based on the 64 codon sequences and the 4 nucleotide sequences. By comparing to independent component analysis (ICA), we can differentiate between the exon and intron regions, and how they are correlated in terms of the square magnitudes of STFTs. The experiment is done on the gene F56F11.4 in the chromosome III of C. elegans. For this data, the nucleotide based PCA identifies the exon and intron regions clearly. The codon based PCA reveals a weak internal structure in some exon regions, but not the others. The result of ICA shows that the nucleotides thymine (T) and guanine (G) have almost all the information of the exon and intron regions for this data. We hypothesize the existence of complex exon structures that deserve more detailed analysis.

Key words : Principal Component Analysis, Exon, Short-Time Fourier Transform, Codon, Nucleotide, Independent Component Analysis

1. Introduction

A large number of DNA sequences are generated from the genome sequencing projects, many of their characterizations are still at a crude form. The whole sequencing of many genomes is in the order of billions of characters in length. Much of these data are raw, in the sense that their relationship to protein coding, functionality and sequence structure are highly complex and difficult to be understood as a whole. However, interpreting this information from a genome sequence or even part of the sequence is one of the most exciting challenges facing biologists and information scientists today.

Given a selected part of a DNA sequence that represents certain specific characteristics, a transformed sequence of the data may reveal some recurring characteristics of the sequence. Genomic sequence information is at the basic level discrete in nature because there are only a finite number of nucleotides in the DNA alphabets. To capture the recurring characteristics, we may interpret the DNA sequence as a discrete-time sequence that can be studied using techniques from the field of digital signal processing such as short-time Fourier transform (STFT).

DNA sequences can be regarded as nonstationary signal whose properties vary with time. However, a single discrete Fourier transform (DFT) estimate is not sufficient to describe such signals. Locating protein coding or exon regions in genomic data has been an important application area of these techniques [1-6]. In [5], the

1) Dept. of Statistical Information, Catholic University of Daegu
2) Dept. of Statistics, Korea University

protein coding regions have been detected by using a method to maximize the discriminatory capability based on the STFT of the DNA sequence as compared to random sequence. In this paper, we further analyze the principal component analysis (PCA) in identifying exons in a gene because of their easier interpretations and evaluate their possible internal structures. The PCA is conducted on the short-time Fourier transform (STFT) as described in [5]. It is noted that when applying PCA to STFT frequency slices, the principal STFT retains explicit information about the nonstationary spectral content of a sequence, as well as implicit information necessary for the reconstruction of the sequence [7]. Similar to [7], we identify exon and intron regions using the first principal component (PC) of the STFTs of the nucleotide sequence. We also use the PCs of STFTs from the codon sequence to investigate whether further internal structures in exons exist. By using independent component analysis (ICA) as described in [8], we identify the nucleotides having information on the exon and intron regions, and how they are correlated in terms of the square magnitudes of the STFTs. These methods are illustrated and analyzed using the gene F56F11.4 in the chromosome III of C. elegans.

2. STFT of Base Sequence

Periodic correlations in DNA sequences can be examined using Fourier analysis. For a DNA sequence of length N, assume that we assign the numbers a, t, c, g to the nucleotide characters A, T, C, G, respectively. The resulting numerical sequence is represented as

$$x[n] = a\,u_A[n] + t\,u_T[n] + c\,u_C[n] + g\,u_G[n], \quad n = 0, 1, 2, \Lambda, N-1 \qquad (1)$$

in which $u_A[n]$, $u_T[n]$, $u_C[n]$ and $u_G[n]$ are the binary indicator sequences, which take the value of either 1 or 0 at location n, depending on whether or not the corresponding character exists at location n.

As defined, any three of these four binary indicator sequences are sufficient to determine the DNA character string, since

$$u_A[n] + u_T[n] + u_C[n] + u_G[n] = 1 \quad \text{for all } n. \qquad (2)$$

A proper choice of the values a, t, c and g for a DNA segment can provide potentially useful properties to the numerical sequence $x[n]$.

The main computational tool that we use is the discrete Fourier transform (DFT) of a sequence $x[n]$ of length N. The DFT is itself another sequence $X[k]$ of the same length N, defined as

$$X[k] = \sum_{n=0}^{N-1} x[n]\, e^{-j\frac{2\pi}{N}kn}, \qquad k = 0, 1, \Lambda, N-1 \qquad (3)$$

The sequence $X[k]$ provides a measure of the frequency content at frequency k, which corresponds to an underlying period of $\frac{N}{k}$ samples. It turns out that, except for finite length effects that can be corrected, the square of the magnitude of the DFT is also a scaled version of the DFT of the autocorrelation sequence. From Eqs. (1) and (3) it follows that

$$X[k] = a\,U_A[k] + t\,U_T[k] + c\,U_C[k] + g\,U_G[k], \quad k = 0, 1, \Lambda, N-1. \qquad (4)$$

For DNA character strings based on the raw sequence, the sequences $U_A[k]$, $U_T[k]$, $U_C[k]$ and

$U_G[k]$ provide a four-dimensional representation of the frequency spectrum of the character string. The quantity

$$S[k] = |U_A[k]|^2 + |U_T[k]|^2 + |U_C[k]|^2 + |U_G[k]|^2 \tag{5}$$

has been used as a measure of the total spectral content of the DNA character string at frequency k. From Eqs. (2) and (3), it follows that

$$U_A[k] + U_T[k] + U_C[k] + U_G[k] = \begin{cases} 0, & k \neq 0 \\ N, & k = 0. \end{cases} \tag{6}$$

Therefore, we can reduce the dimensionality of the frequency spectrum representation from four to three, for example, by ignoring one of the four frequency components.

DNA sequence can be viewed as nonstationary signal whose properties vary with time and analyzed from this perspective. A single DFT estimate is not sufficient to describe such signals, and as a result, the use of the STFT is proposed. We can apply a sliding window of small length to a sequence, resulting in a sequence of DFTs', each providing a localized measure of the frequency content. This is known as the STFT that is defined as follows.

$$X[n,k] = \sum_{m=0}^{L-1} x[n+m]\, w[m]\, e^{-j\frac{2\pi}{N}km}, \qquad k = 0, 1, \Lambda\ , N-1 \tag{7}$$

where $w[m]$ is a window sequence. Then $X[n,k]$ is the DFT of the windowed sequence $x[n+m]\, w[m]$. Here, we suppose that the window has length L with samples beginning at $m = 0$. The primary purpose of the window in the STFT is to limit the extent of the sequence to be transformed so that the spectral characteristics are reasonably stationary over the duration of the window. A typical window for spectrum analysis tapers to zero so as to select only a portion of the signal for analysis. Similar to [5], we here use the STFT of a DNA sequence using a Hamming window of given length (of 351). The Hamming window is defined by

$$w(m) = 0.54 + 0.46 \cos\left(\frac{2\pi m}{2L+1}\right), \qquad m = 0, 1, \Lambda\ , L \tag{8}$$

If the window length L is too long, the signal properties may change too much across the window. If the window is too short, resolution of narrowband components will be sacrificed. This is typical of the trade-off between frequency resolution and time resolution that is required in the analysis of nonstationary signals.

In the STFT, the one-dimensional sequence $x[n]$, a function of a single discrete variable, is converted into a two-dimensional function of the time variable n, which is discrete, and the frequency variable k, which is also discrete. Another way to see that the STFT can be sampled in the time dimension is to recall that for fixed k, the STFT is a one-dimensional sequence that is the output of a bandpass filter with frequency response. If we sample $X[n,k]$ at N equally spaced frequencies k, then we can recover the original sequence $x[n]$ from the STFT. It can be reconstructed if $X[n,k]$ is sampled in the time dimension as well.

The display of the magnitude of the STFT is called a spectrogram. The frequency $k = \frac{L}{3}$ corresponds to a period of three samples, equal to the length of each codon. It is known that the spectrum of protein coding DNA typically has a peak at that frequency. This property has been used to design a gene prediction algorithm. If we define the following normalized DFT coefficients at frequency $k = \frac{L}{3}$:

$$W = \frac{1}{N} X[n, \frac{L}{3}]$$
$$A = \frac{1}{N} U_A[n, \frac{L}{3}],\ T = \frac{1}{N} U_T[n, \frac{L}{3}],\ C = \frac{1}{N} U_C[n, \frac{L}{3}],\ G = \frac{1}{N} U_G[n, \frac{L}{3}] \quad (9)$$

Then, from Eq. (4), with $k = \frac{L}{3}$, it follows that

$$W = a A + t T + c C + g G. \quad (10)$$

In other words, for each DNA segment of length L (where L is a multiple of 3), and for each choice of the parameters a, t, c and g, there corresponds a complex number $W = a A + t T + c C + g G$. It has been found that for properly chosen values of a, t, c and g, W can be an accurate predictor of a protein coding segment, but furthermore, the reading frame of the segment. The latter information is coming from the phase of W, or $\Theta = \arg\{W\}$.

For each sequence segment, there corresponds a set of complex numbers A, T, C and G which satisfies $A + T + C + G = 0$. The quantity W is a complex random variable and its properties depend on the particular choice of the parameters a, t, c and g. Parameters which maximize the discriminatory capability between exon regions (with corresponding random variables A, T, C and G) and random sequence can be used to find exon regions in a gene. In this paper, we further analyze this capability here using PCA analysis with a more direct interpretation.

3. Experiment using F56F11.4 Gene Sequence Data

The experiment is to analyze the F56F11.4 gene sequence from C. elegans. The gene is known to be divided into five exon regions with intron regions between them. Our goal is to identify these regions from the sequence signals of the STFT transform using PCA and ICA analysis. When these analyses identify these regions, we further evaluate the internal structures of the identified exon regions. Since these methods reflect a meaningful interpretation of these regions, the existence of these internal exon structures can be evaluated. In the experiment, we calculate the STFT at frequency $k = \frac{L}{3}$, $L = 351$.

3.1 Description of the data

Caenorhabditis elegans is a small (about 1 mm long) soil nematode found in temperate regions. In the 1960's Sydney Brenner began using it to study the genetics of development and neurobiology. Since then the community of C. elegans researchers has expanded to over a thousand. Around the world many hundreds of scientists are working full time investigating the biology of C. elegans. Between October, 1994 and January, 1995, 73 scientific articles about this creature appeared in international science journals.

In this paper, we deal with the gene F56F11.4 containing 8,000 nucleotides starting from location 7021 in the chromosome III of C. elegans (accession number AF099922). The gene contains five exons which are

known to be at the following positions, relative to 7021. The relative exon locations are described in Table 1.

Table 1. Locations and reading frames of the five exons of gene F56F11.4

Relative position	Exon length	Reading frame
929-1135	207	2
2528-2857	330	2
4114-4377	264	1
5465-5644	180	2
7255-7605	351	1

3.2 Analysis results of PCA based on the nucleotide sequence

We will here study how well the PCA identifies the exon and intron regions, and then use the analysis to evaluate the internal exon structures. First, we take into consideration the corresponding STFT in Eq. (9) at frequency $k = \frac{L}{3}$, $L = 351$ for each nucleotide. Second, we put them together so as to make 4-dimensional data at each location. Third, we employ the PCA to these 4-dimensional complex data in order to study how well the PCA identifies the exon and intron regions. Last, we compare this method with the STFT with the choice of the parameters $a = 0.10 + 0.12j$, $t = -0.30 - 0.20j$, $c = 0$ and $g = 0.45 - 0.19j$. The analysis results are shown in Fig. 1.

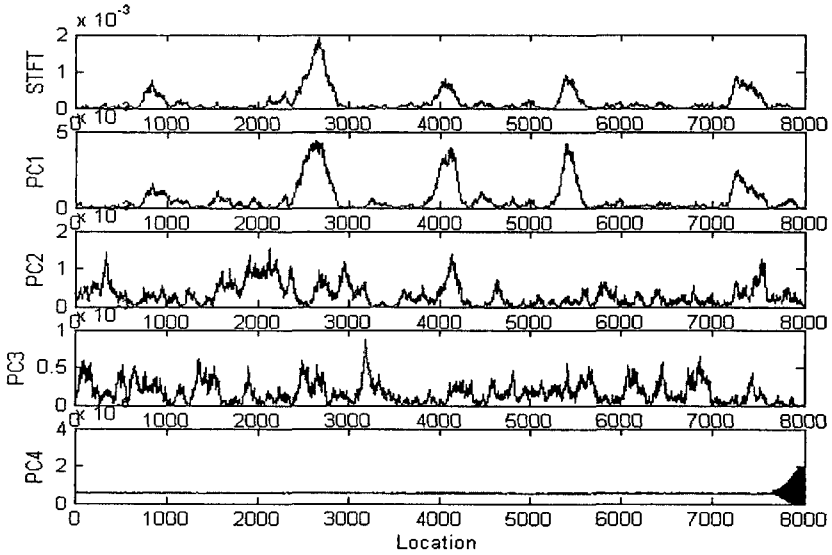

Fig. 1. Square magnitudes of the STFT and the 4 PCs

We here describe the square magnitudes of STFT and the 4 principal components (PCs). In Fig. 1, the first plot is for $|aA + tT + cC + gG|^2$, and the next four plots are for the square magnitudes of PC1, PC2, PC3 and PC4, respectively.

References

D. Anastassiou(2000), Frequency-domain analysis of biomolecular sequences. Bioinformatics 16 (2000) 1073-1081.

D. Beyerbach, H. Nawab(1991), Principal components analysis of the short-time Fourier transform, Proc. IEEE ICASSP, 1725-1728.

E. Bingham, A. Hyvrinen(2000), A fast fixed-point algorithm for independent component analysis of complex-valued signals. Int. J. of Neural Systems, 10 (1), 1-8.

J.W. Fickett(1982), Recognition of protein coding regions in DNA sequences, Nucleic Acids Research, 10, 5303-5318.

S.K. Mitra(2000), Digital Signal Processing: A Computer-Based Approach, 2ndedn, McGraw-Hill, New York.

S.L. Salzberg(1995), Locating protein coding regions in human DNA using a decision tree algorithm, Journal of Computational Biology 2 (3), 473-485.

B.D. Silverman, R. Linsker(1986), A measure of DNA periodicity, J. Theor. Biol. 118, 295-300.

E.E. Snyder, G.D. Stormo(1995), Identification of protein coding regions in genomic DNA,Journal of Molecular Biology 248, 1-18.

S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, R. Ramaswamy(1997), Prediction of probable genes by Fourier analysis of genomic sequences, CABIOS 113, 263-270.

P.P. Vaidyanathan, B. Yoon(2002), Gene and exon prediction using allpass-based filters. Workshop on Genomic Signal Processing and Statistics (GENSIPS), Raleigh, NC, 2002.