Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

On relationship of Z-curve and Fourier approaches for DNA coding sequence classification

Ngai-Fong Law*, Kin-On Cheng and Wan-Chi Siu

Centre for Multimedia Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong; N. F. Law * - E-mail: ennflaw@polyu.edu.hk; Phone: +852 2766 4746; Fax: +852 2362 8439; * Corresponding author

received October 20, 2006; accepted November 02, 2006; published online November 14, 2006

Abstract:

Z-curve features are one of the popular features used in exon/intron classification. We showed that although both Z-curve and Fourier approaches are based on detecting 3-periodicity in coding regions, there are significant differences in their spectral formulation. From the spectral formulation of the Z-curve, we obtained three modified sequences that characterize different biological properties. Spectral analysis on the modified sequences showed a much more prominent 3-periodicity peak in coding regions than the Fourier approach. For long sequences, prominent peaks at 2Π/3 are observed at coding regions, whereas for short sequences, clearly discernible peaks are still visible. Better classification can be obtained using spectral features derived from the modified sequences.

Keywords: Z-Curve approach; FT analysis; DNA Sequence; coding region; spectral analysis

Background:

A DNA sequence is a long sequence consisting of four types of nucleotides: Adenine (A), Guanine (G), Thymine (T) and Cytosine (C). An important problem for sequence analysis is to distinguish coding (exons) and non-coding (introns and intergenic spaces) regions in a sequence. Sequence features exploiting properties such as codon usage bias, base compositional bias between codon positions, periodicity in base occurrence in coding regions **[1, 2]** have been proposed for characterizing coding/non-coding regions. The 3-periodicity property of coding regions is particularly interesting and has been studied intensely. A natural choice for detecting such periodicity is the Fourier Transform (FT).

The Z-curve features **[3]** and the FT approach **[4-6]** are both concerned with detecting the 3-periodicity property of coding sequences and are implicitly related. However, there is no theoretical study of the relationship between the two approaches. In this paper, we give a theoretical analysis that reveals the relationships between the two and show that there are significant differences among them. In particular: (1) we provide a theoretical study of the relationship between the two approaches; (2) we provide a justification for the empirical observation that Z-curve approach generally have better performance than FT approach, especially for shorter sequences; and finally, (3) we propose a modification of the basic FT approach based on a new numerical sequence representation derived from Z-curve that preserves biological significance.

Methodology:

Spectral-Based DNA Sequence Analysis

A DNA sequence of length N can be written as $S = S_0 S_1 ... S_{N-1}$ where $S_i \in \{A,T,G,C\}$. Typically, the DNA sequence is rewritten as **[5]**,

$$x[n] = au_A[n] + gu_G[n] + tu_T[n] + cu_C[n] \quad (1)$$

where $u_A[n]$, $u_T[n]$, $u_C[n]$ and $u_G[n]$ are binary indicator sequences and $\{a,t,c,g\}$ are weightings associated with the corresponding binary sequences. The binary indicator sequences take the value of 1 or 0 at location *n*, depending upon whether the corresponding character exists at *n*.

The goal of performing spectral analysis on DNA sequences is to highlight sequence structure and frequency components that may be present. Discrete Fourier transform (DFT) can be applied to the numerical sequence to analyze its spectral features. In particular, the power spectrum can be formed as

$$\tilde{X}[k] = \sum_{j\in\{A,G,T,C\}} \left|\tilde{U}_j[k]\right|^2 \quad (2)$$

where $\tilde{U}[k] = \sum_{n=0}^{N-1} u[n] e^{-j\frac{2\pi kn}{N}}$. The spectral approach for DNA sequence analysis relies on the assumption that the spectrum for coding region is different from that for non-coding region due to codon usage bias. In particular, a coding region is identified if a peak at frequency 2Π/3 is observed. However, the magnitude of the peak varies greatly. To increase the discriminating power, one can adjust the four weights, $\{a,g,t,c\}$, in (1). For example, Anastassiou **[5]** obtained the weightings through an optimization process which maximizes the differences between the spectra formed from exons and introns in a set of "training" sequences.

Z-Curve Approach

The Z-curve approach **[3]** extracts features directly from the character-based DNA sequence. In particular, statistical information about the cumulative frequencies of the occurrence of individual nucleotide is used. Let the frequencies of bases A, C, G and T at positions 0, 3, 6,; 1, 4, 7, … and 2, 5, 8, … respectively be $A_0, C_0, G_0, T_0; A_1, C_1, G_1, T_1; A_2, C_2, G_2, T_2$; the nine features in the Z-curve approach are then defined as

$$\begin{aligned} f_{3i} &= (A_i + G_i) - (C_i + T_i) \\ f_{3i+1} &= (A_i + C_i) - (G_i + T_i) \\ f_{3i+2} &= (A_i + T_i) - (C_i + G_i), \qquad i = 0,1,2 \end{aligned} \tag{3}$$

The biological interpretation of the above three measures are as follows **[3]**: component f_{3i} displays the distribution of bases of the purine (A or G) and pyrimidine (C or T) types along the sequence. Component f_{3i+1} displays the distribution of the bases of amino (A or C) and keto (G or T) types. Component f_{3i+2} displays the distribution of the bases of the weak H-bond (A or T) and strong H-bond (G or C) types. These nine values form a feature vector which helps to distinguish the coding from non-coding regions. For example, a neural network can be employed or a Fisher discriminate analysis can be used to perform the classification. **[3]**

Relationship between Z-Curve & Spectral Approaches

Both the spectral approach and the Z-curve exploit the three-periodicity in the coding region. To elucidate the relationship between the two, we studied the Z-curve features from a signal processing perspective. Using the binary indicator sequence $u_b[n]$, the cumulative frequencies A_i, C_i, G_i, T_i can be written as,

$$b_i = \frac{1}{N}\sum_{n=0}^{N-1} l_i[n]u_b[n] \qquad b \in \{A,G,T,C\}, i = 0,1,2 \tag{4}$$

where l_i captures information relating to the nucleotide position and is defined as,

$$l_i = \sum_{m=0}^{\frac{N}{3}-1} \delta[n-3m-i] \qquad i = 0,1,2 \tag{5}$$

Using (4) and (5), (3) can be written as,

$$f_{3i+i'} = \frac{1}{N}\sum_{n=0}^{N-1} s_{i'}[n]l_i[n] \qquad i' = 0,1,2 \tag{6}$$

where $s_0[n], s_1[n]$ and $s_2[n]$ are the modified sequences and are formed from the DNA sequence $x[n]$ with {a,g,t,c} equals to {1,1,-1,-1}, {1,-1,-1,1} and {1,-1,1,-1}, respectively. Using the Parvesal's theorem, (6) can be written in the frequency domain as

$$f_{3i+i'} = \frac{1}{N^2}\sum_{k=0}^{N-1} \tilde{S}_{i'}[k]\tilde{L}_i^*[k] \tag{7}$$

where $\tilde{S}_{i'}[k]$ is the N-point DFT of the modified sequences $s_{i'}[n]$, $\tilde{L}_i[k]$ is the N-point DFT of sequence l_i defined in (5) and * denotes complex conjugate. Note that the DFT of sequence l_i can be written as a sum of delta functions,

$$\tilde{L}_i[k] = \frac{N}{3}\sum_{m=0}^{2} \delta\left[k - \frac{Nm}{3}\right] e^{-j\frac{2\pi ik}{N}} \tag{8}$$

Substituting (8) into (7) gives,

$$f_{3i+i'} = \frac{1}{3N}\sum_{m=0}^{2} \tilde{S}_{i'}\left[\frac{Nm}{3}\right] e^{j\frac{2\pi im}{3}} \tag{9}$$

Using the conjugate property of a real sequence, (9) can be written as

$$f_{3i+i'} = \frac{1}{3N}\left\{\tilde{S}_{i'}[0] + 2\,\mathrm{Real}\left[\tilde{S}_{i'}\left[\frac{N}{3}\right] e^{j\frac{2\pi i}{3}}\right]\right\} \tag{10}$$

Eq (10) shows the relationship between the Z-curve features $f_{3i+i'}$ and the spectra of the three modified sequences $s_{i'}[n]$ at frequency N/3. It clearly shows that the Z-curve features measure different compositions of the nucleotides along the sequence as well as any 3-periodicity present in the coding region. For example, the DC value $\tilde{S}_0[0]$ measures the difference between the distribution of the bases of purine and pyrimidine types, $\tilde{S}_1[0]$ measures the difference between the distribution of the bases of amino and keto types, and $\tilde{S}_2[0]$ measures the difference between the distribution of the bases of the weak H-bond and strong H-bond. The term $\tilde{S}_{i'}[N/3]$ implies a sampling at every third position and detects the 3-periodicity characteristic. If the magnitude of $\tilde{S}_{i'}[N/3]$ is large, a peak is observed in which a coding region is identified.

Comparative Analysis

Both the FT approach in (2) and the Z-curve approach ((9) or (10)) attempt to measure the 3-periodicity in the DNA sequence. Nevertheless, there are significant differences between them.

Although $s_0 = u_A + u_G - u_T - u_C$, $|S_0|^2 \neq |U_A|^2 + |U_G|^2 - |U_T|^2 - |U_C|^2$. Thus, the weighting in the FT approach is different from the weighting used in the Z-curve approach. The former considers each spectrum independently as $\sum_{j\in\{A,G,T,C\}} w_i |U_j|^2$ while the Z-curve approach considers the spectra of the modified sequences. The periodicity assumption is also different between the FT approach and the Z-curve approach. In the Z-curve approach, the periodicity assumption applies with regards to the biological properties and the nucleotide positions induced by the different base combination. In contrast, the periodicity assumption in the FT approach is made regardless of the biological properties. It simply sums up the spectra of different nucleotide indicator sequences independently. To demonstrate, lets consider an artificial sequence {T, A, G, C, G, A}. In the FT approach, this gives rise to four binary indicator sequences, {0, 1, 0, 0, 0, 1} (A), {0, 0, 1, 0, 1, 0} (G), {1, 0, 0, 0, 0, 0} (T) and {0, 0, 0, 1, 0, 0} (C). Periodicity cannot be observed in any sequence. In the Z-curve approach, the modified sequence $s_0[n]$ is {-1, 1, 1, -1, 1, 1}, which shows strong 3-periodicity. Finally, three modified sequences which characterize different biological properties are considered in the Z-curve approach whereas only the original

sequence is considered in the FT approach. Hence, the FT approach considers only one spike at 2Π/3 for classification whereas the Z-curve approach considers both the DC value and the value at 2Π/3 of three modified sequences.

In view of the above analysis, we proposed to apply the FT to the modified sequences $s_0[n], s_1[n]$ and $s_2[n]$. The power spectra of the modified sequences are first formed. The three DC values and the three values at 2Π/3 can then be used for sequence classification. These DC and 2Π/3 features are in fact closely related to the nine Z-curve features as seen in (10), and they carry similar biological interpretation as the Z-curve features.

Results & Discussion:

We used human exon and intron datasets downloaded at http://www.ncbi.nlm.nih.gov/ for testing. In our first result, we chose exons and introns with a length approximately equal to 1500. Fig. 1 shows $\tilde{X}[k]$ in the FT approach and the spectrum for $s_2[n]$. It can be seen that both approaches can detect the 3-periodicity in the coding regions as peaks are observed at *k*=501 which corresponds to the 2Π/3 frequency. Fig. 2 shows results for another exon with GenBank accession number "AX136319". In the FT approach, the peak cannot be easily identified. In contrast, the spectrum of $s_2[n]$ clearly shows the peak at 2Π/3 (*k*=411). As discussed in Section IV, this is due to the fact that the periodicity assumption is made with respect to the biological property embedded in the DNA sequence.

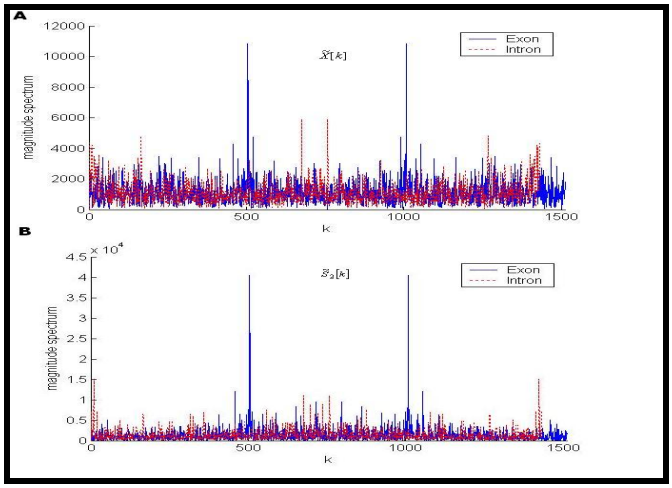

Figure 1: Magnitude Spectrum in both coding (exons) and non-coding (introns) regions

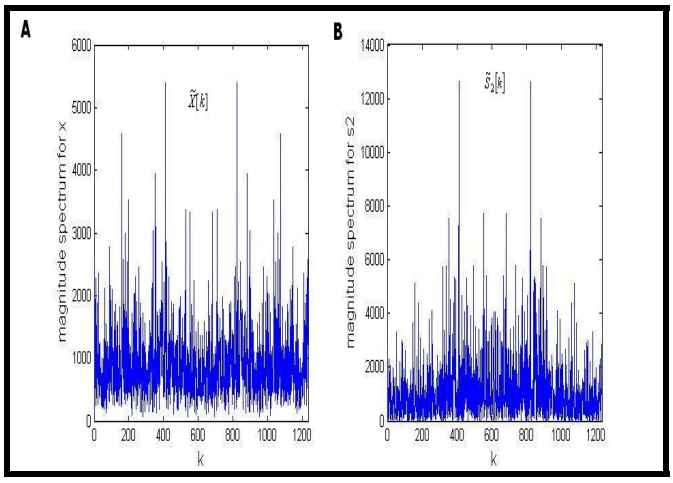

Figure 2: Magnitude Spectrum for 'AX136319'

Recognizing human exons is sometimes a very challenging problem as human exons can be very short in length (137 bp in average). Exon sequences with GenBank accession numbers "AB061839" and "AB050050" are chosen for testing. The first sequence has 123 bp while the second sequence has 127 bp. Results for these sequences are shown respectively in Fig. 3 and Fig. 4. Due to the short length of the

exon sequences, peaks are not observed at 2Π/3 for both sequences for $\tilde{X}[k]$. In contrast, peaks are observed at 2Π/3 (k=41 and k=43 respectively for Fig. 3 and Fig. (4) for the spectrum of the modified sequences $s_i[n]$.

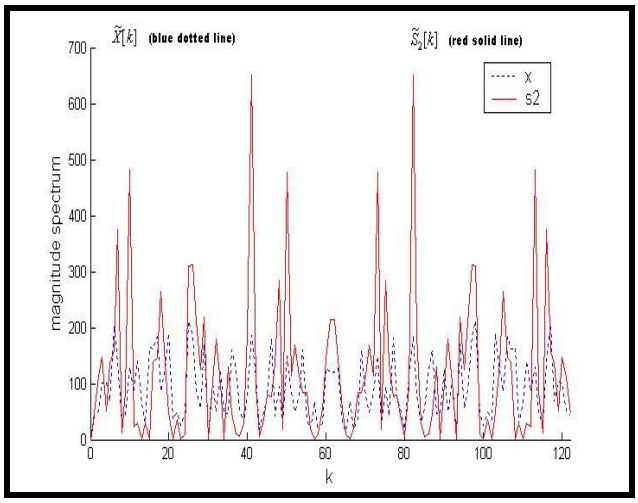

Figure 3: Magnitude spectrum for 'AB061839'. No discernible peaks can be observed for blue dotted line at 2Π/3 (k=41).

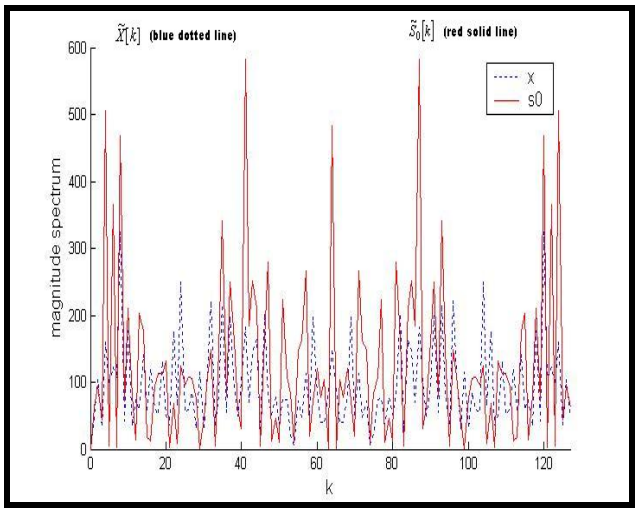

Figure 4: Magnitude spectrum for 'AB050050'. No discernible peaks can be observed for blue dotted line at 2Π/3 (k=43).

In a further experiment, we extracted two different datasets. **[7]** These datasets consist of 6000 Yeast ORFs and 6000 Yeast No Feature sequences, 1500 human exons and 1500 introns whose length is less than 140bp. We performed classification experiments for both the FT approach and our proposed approach. In the FT approach, the value at 2Π/3 is extracted as the feature. In our proposed approach, the three power spectra of $s_0[n]$, $s_1[n]$ and $s_2[n]$ are firstly obtained. Then the features for classification are the values at 2Π/3 and the DC values for these three spectra.

Classification experiments using these selected features were then performed using the k-nearest-neighbor classifier as in Wu *et al.*, **[8]** Table I summarizes the results. Note that sensitivity is defined as the proportion of coding sequences that have been correctly classified as coding while specificity is the proportion of non-coding sequences that have been correctly classified as non-coding. From Table 1, we see that for human sequences, low specificity is observed for the FT approach. This implies that many non-coding sequences are wrongly classified as coding sequences. However, using the proposed approach, the performance is greatly improved. For Yeast sequences, both sensitivity and specificity are increased by using the proposed features.

	Yeast	Human
FFT approach		
Sensitivity	0.8580	0.8627
Specificity	0.8922	0.2873
Average	0.8751	0.5750
Proposed approach		
Sensitivity	0.8607	0.7607
Specificity	0.9558	0.8413
Average	0.9083	0.8010

Table 1: Classification results of coding and non-coding sequences

Conclusion:

Z-curve features are one of the popular features used for DNA sequence classification and they are closely related to a FT spectral analysis of the sequence for 3-periodicity. In this paper we gave a theoretical study of the relationship between the Z-curve and the FT approach. Our analysis showed that there are significant differences in the spectral interpretation between the two. We discussed the implications of these differences for shorter sequences. Moreover, we showed that the three modified sequences obtained from the spectral reformulation of the Z-curve approach characterize different biological properties and are useful for coding region prediction. In particular, the 3-periodicity is much more prominent in the modified sequences. As a result of our analysis, we proposed to apply spectral analysis to the three modified sequences to better capture the 3-periodicity property embedded in the coding region of a DNA sequence and verified this experimentally.

Acknowledgment:

This work is supported by RGC Grant PolyU 5210/04E, the project A-PA2P and the Centre for Multimedia Signal Processing (A452), the Hong Kong Polytechnic University. The authors have no conflict of interest in this work.

References:

[01] R. Staden & A. D. McLachlan, *Nucleic Acids Res.*, 10:141 (1982) [PMID: 7063399]

[02] J. W. Fickett, *Nucleic Acids Res*., 10:5303 (1982) [PMID: 7145702]

[03] C. T. Zhang & J. Wang, *Nucleic Acids Res*., 28:2804 (2000) [PMID: 10908339]

[04] S. Tiwari *et al.*, *Comput Appl Biosci*., 13:263 (1997) [PMID: 9183531]

[05] D. Anastassiou, *Bioinformatics*, 16:1073 (2000) [PMID: 11159326]

[06] B. Isaac, *et al.*, *Bioinformatics*, 18:196 (2002) [PMID: 11836230]

[07] A. W. C. Liew, *et al*., *Int. J. of Bioinformatics Res. and Applications*, 1:181 (2005)

[08] Y. Wu, *et al*., *Phys. Rev. E.*, 67:061916 (2003) [PMID: 16241270]

Edited by P. Kangueane

Citation: Law *et al.*, Bioinformation 1(7): 242-246 (2006)