

Spectral Approaches for DNA Sequence Classification

K. O. Cheng, N. F. Law and W. C. Siu

Abstract-- Z-curve features are one of the popular features used in DNA sequence classification. Here, we studied the Z-curve features from a signal processing point of view. In particular, the Z-curve features are re-interpreted through a spectral formulation. Our analysis showed that there are significant differences in the spectral interpretation between the Z-curve formulation and the FFT (Fast Fourier Transform) approach. From the spectral formulation of the Z-curve approach, we obtained three modified sequences that characterize different biological properties which are useful for coding region prediction. Spectral analysis on the modified sequences showed a much more prominent three-periodicity property in coding regions than using the FFT approach. Our experiments indicated that for long sequences, prominent peaks at $2\pi/3$ are observed at coding regions. For short sequences, peaks can still be observed at coding regions. We also obtained good classification performance using the spectral features derived from the three modified sequences.

Index Terms—Z-curve approach, Fourier approach, DNA sequence, coding region, spectral analysis

I. INTRODUCTION

A DNA sequence is a long sequence consisting of four types of nucleotides: Adenine (A), Guanine (G), Thymine (T) and Cytosine (C). Studying the structure characteristics of this sequence is one of the most important research problems in Bioinformatics. Analytical tools for sequence analysis rely on finding features that can distinguish coding (exons) from non-coding (introns and intergenic spaces) regions. Sequence features such as codon usage measure, base compositional bias between codon positions, periodicity in base occurrence and Z-curve features [1]-[4] provide a statistical means for characterizing these regions in the sequence. Recently, signal processing (SP) techniques have been proposed to study the DNA sequence [5]. The use of SP technique relies on finding a mapping that transforms the character sequences to numerical

sequences. A good numerical representation must be able to capture all significant properties of the biological reality without introducing any spurious effects. In this paper, we reformulated the Z-curve features that have been used in DNA sequence study from a SP point of view. This analysis reveals relationships between the popular Z-curve features and the FFT (Fast Fourier Transform) approach for DNA sequences classification. Our analysis shows that although both approaches are based on detecting the three periodicity in the coding sequence, there are significant differences among them. From the analysis, we also derive numerical sequence representations that preserve biological significance.

II. SPECTRAL-BASED SEQUENCE ANALYSIS

A DNA sequence of length N can be written as $S = S_0 S_1 \dots S_{N-1}$ 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 either 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

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

where $\tilde{U}_i[k] = \sum_{n=0}^{N-1} u_i[n] e^{-j \frac{2\pi kn}{N}}$. 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. In particular, a coding region is identified if a peak at frequency equals to $2\pi/3$ is observed [4]-[7]. Otherwise, a non-coding region is assumed. 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). Different choices of these values would give different numerical sequences which would affect the presence/absence of the peak at $2\pi/3$ [5]-[6]. For example, [6] chooses the weightings as projections from a tetrahedral representation while [5] obtained the weightings

K. O. Cheng is with Centre for Multimedia Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University (corresponding author to provide phone: 852-2766 6201; fax: 852-2362 8439; e-mail: k.o.cheng@polyu.edu.hk).

N. F. Law is with Centre for Multimedia Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University (phone: 852-2766 4746; fax: 852-2362 8439; e-mail: ennflaw@polyu.edu.hk).

W. C. Siu is with Centre for Multimedia Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University (phone: 852-2766 6229; fax: 852-2362 8439; e-mail: enwcsiu@polyu.edu.hk).

through an optimization process which tries to maximize the differences between the spectra formed from exons and introns in a set of “training” sequences.

III. Z-CURVE APPROACH

The Z-curve approach [1] extracts features directly from the character-based DNA sequence. In particular, statistical information about the cumulative frequencies of the occurring 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

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

The biological interpretation of the above three measures are as follows [1]: 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 the non-coding regions. For example, a neural network can be employed or a Fisher discriminate analysis can be used to perform the classification [1].

IV. RELATIONSHIP BETWEEN Z-CURVE AND SPECTRAL APPROACH

Both the spectral approach and the Z-curve exploit the three-periodicity in the coding region. To elucidate the relationship between the two approaches, we studied the Z-curve features from a SP point of view. Using the binary indicator sequences $u_b[n]$, the cumulative frequencies A_i, C_i, G_i, T_i can be rewritten as,

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

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

$$l_i = \sum_{m=0}^{\frac{N-1}{3}} \delta[n - 3m - i] \quad i \in \{0, 1, 2\} \quad (5)$$

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

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

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 [8], the sum in the RHS of (6) can be rewritten in the frequency domain as

$$f_{3i+i'} = \frac{1}{N^2} \sum_{k=0}^{N-1} \tilde{s}_{i'}[k] \tilde{l}_i^*[k] \quad (7)$$

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 rewritten as a sum of delta functions,

$$\tilde{l}_i[k] = \frac{N}{3} \sum_{m=0}^2 \delta\left[k - \frac{Nm}{3}\right] e^{-j\frac{2\pi mk}{N}} \quad (8)$$

Substituting (8) into (7) gives,

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

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

$$f_{3i+i'} = \frac{1}{3N} \left\{ \tilde{s}_{i'}[0] + 2\text{Real} \left[\tilde{s}_{i'}\left[\frac{N}{3}\right] e^{j\frac{2\pi im}{3}} \right] \right\} \quad (10)$$

Equation (10) shows the relationship between the Z-curve features $f_{3i+i'}$ and the spectra of the three modified sequences.

The DC value $\tilde{s}_{i'}[0]$ measures different compositions of the nucleotides along the sequence. In particular, $\tilde{s}_0[0]$ measures the difference between the distribution of the bases of purine and pyrimidine, $\tilde{s}_1[0]$ 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 $\tilde{s}_{i'}\left[\frac{N}{3}\right]$

relates the three-periodicity characteristics of the modified sequence $s_{i'}[n]$. If $\tilde{s}_{i'}\left[\frac{N}{3}\right]$ is very large, a peak is observed in

which a coding region is identified. If $\tilde{s}_{i'}\left[\frac{N}{3}\right]$ is small, a

non-coding region is found. Using (10), one can relate the Z-curve features and the spectral features. Specifically, it clearly shows that the Z-curve features measure different compositions of the nucleotides along the sequence together with the three-periodicity characteristics of the coding region. If there is strong three-periodicity in the sequence, $f_{3i+i'}$ would be large.

V. COMPARATIVE ANALYSIS

Both the approach in (2) (named FFT approach) and the Z-curve approach ((9) or (10)) capture the three-periodicity in the DNA sequence. However the FFT approach operates on the “complete” sequence, while the Z-curve features are extracted from a “down-sampled by three” version of the modified sequences ((4)). Moreover, the FFT approach considers the value of a single spike at $2\pi/3$ for the classification of coding and non-coding regions. In contrast, the Z-curve features consider both the DC value and the value at $2\pi/3$.

Although $s_0 = u_A + u_G - u_T - u_C$, it can be seen that $|s_0|^2 \neq |u_A|^2 + |u_G|^2 - |u_T|^2 - |u_C|^2$. There are two implications for this. First, applying a weighting in the spectrum domain is fundamentally different from the weighting used in the Z-curve approach. The former considers each spectrum independently as $\sum_{j \in \{A, G, T, C\}} w_j |U_j|^2$ while the Z-curve approach considers the spectra of the modified sequences. Second, the periodicity assumption in the FFT approach and the Z-curve approach are somewhat different. The periodicity assumption in the Z-curve applies with regards to the biological properties and the nucleotide positions. In contrast, the periodicity assumption in the FFT approach is made regardless of the biological properties. It simply sums up the spectra of different nucleotide sequences independently. To demonstrate the idea, let us consider an artificial sequence {T, A, G, C, G, A}. In the FFT 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 sequences. In the Z-curve approach, the modified sequence $s_0[n]$ is $\{-1, 1, 1, -1, 1, 1\}$, which shows the three-periodicity. Moreover, in the Z-curve approach, three sequences are formed which characterize different biological properties. In contrast, only one sequence is considered in the FFT approach.

In view of the differences between the FFT approach and the Z-curve features, we proposed to apply the FFT to the three modified sequences, i.e., $s_0[n]$, $s_1[n]$ and $s_2[n]$. The power spectra of the three modified sequences are first formed. The three DC values and the three values at $2\pi/3$ can then be used for sequence classification. These DC and $2\pi/3$ values features are in fact closely related to the nine Z-curve features as seen in (10). Thus, these six features carry similar biological interpretation as the Z-curve features.

VI. RESULTS

We used exon and intron datasets downloaded at <http://www.ncbi.nlm.nih.gov/> for testing. Various lengths have been chosen. In our first result, we chose exons and introns with a length approximately equal to 1500. Fig. 1 shows $\tilde{X}[k]$ in the FFT approach and the spectrum for $s_2[n]$. It can be seen that both approaches can detect the three-periodicity in the coding regions as peaks are observed at $k=501$ which corresponds to the $2\pi/3$ frequency. Fig. 2 shows results for another exon with GenBank accession number ‘‘AX136319’’. In the FFT approach, the peak cannot be easily identified. In contrast, the spectrum of $s_2[n]$ clearly shows the peak at $2\pi/3$ ($k=411$). 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.

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 $2\pi/3$ for both sequences for $\tilde{X}[k]$. In contrast, peaks are observed at $2\pi/3$ ($k=41$ and $k=43$ respectively for Fig. 3 and Fig. 4) for the spectrum of the modified sequences $s_i[n]$.

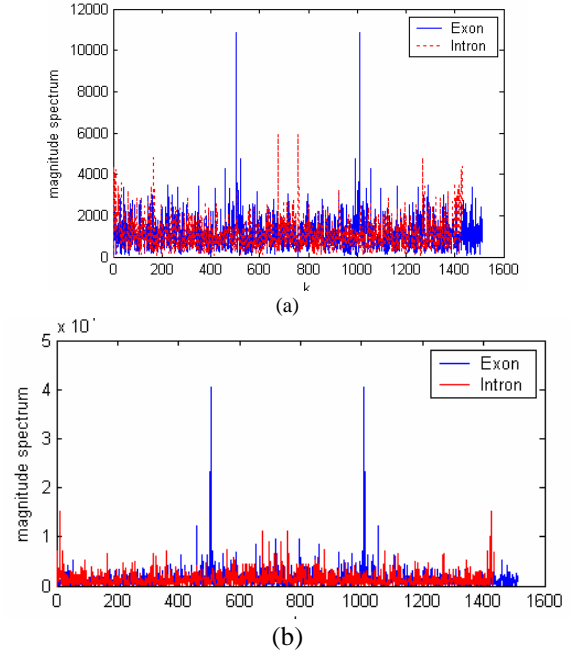


Fig. 1. (a) $\tilde{X}[k]$ and (b) magnitude spectrum for $s_2[n]$ in both coding (exons) and non-coding (introns) regions.

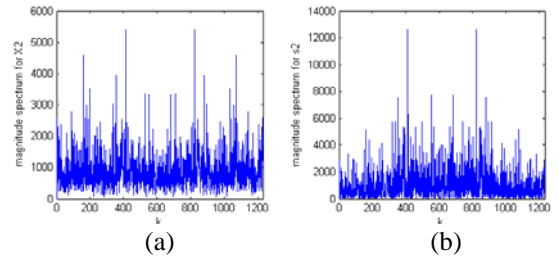


Fig. 2. (a) $\tilde{X}[k]$ and (b) magnitude spectrum for $s_2[n]$ in the sequence ‘‘AX136319’’.

In a further experiment, we extracted two different dataset [9]. These dataset 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 FFT approach and our proposed approach. In the FFT approach, the value at $2\pi/3$ 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 DC values and $2\pi/3$ values from these three spectra.

Classification experiments using these selected features were then performed using the k -nearest-neighbor classifier as in [10]. 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 [11]. From Table I, we see that for human sequences, low specificity is observed for the FFT 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.

TABLE I. CLASSIFICATION RESULTS.

	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

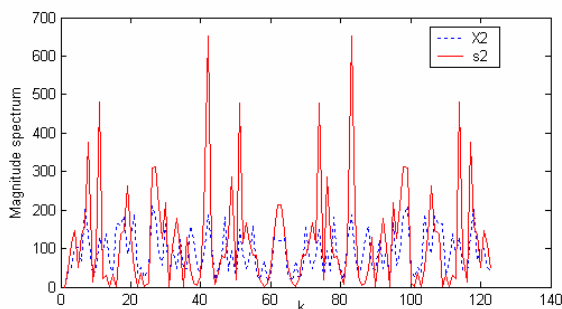


Fig. 3. $\tilde{X}[k]$ (blue dotted line) and the magnitude spectrum for $s_2[n]$ (red solid line) in the sequence 'AB061839'.

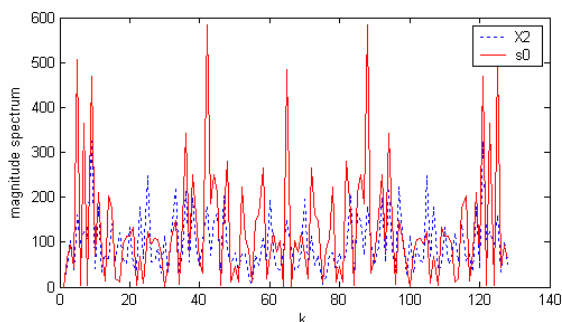


Fig. 4. $\tilde{X}[k]$ (blue dotted line) and the magnitude spectrum for $s_0[n]$ (red solid line) in the sequence 'AB050050'.

VII. CONCLUSIONS

Z-curve features are one of the popular features used for DNA sequence classification. We have reformulated the Z-curve features from a signal processing point of view and gave a spectral interpretation of these features. We show that there are significant differences in the spectral interpretation between the Z-curve formulation and the FFT approach. Moreover, 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. We propose to apply spectral analysis to the three modified sequences to better capture the three-periodicity property embedded in the coding region of the original DNA sequence. The value at $2\pi/3$ and the DC values from the three spectra can then be used as features for classification. Using our proposed approach, we have obtained good classification performance. In particular, a significant improvement is observed for human datasets.

ACKNOWLEDGMENT

This work is supported by RGC Grant PolyU 5210/04E, the project A-PA2P and the Centre for Multimedia Signal Processing, the Hong Kong Polytechnic University.

REFERENCES

- [1] C.T. Zhang and Ju Wang, "Recognition of Protein Coding Genes in the Yeast Genome at Better Than 95% Accuracy based on the Z Curve", *Nucleic Acids Research*, Vol. 28, pp. 2804-2814, 2000.
- [2] R. Staden and A.D. McLachlan, "Codon Preference and its Use in Identifying Protein Coding Regions in Long DNA Sequences", *Nucleic Acids Research*, Vol. 10, Number 1, pp. 141-156, 1982
- [3] J.W. Fickett, "Recognition of Protein Coding Regions in DNA Sequences", *Nucleic Acids Res.*, Vol. 10, pp 5303 - 5318, 1982
- [4] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya and R. Ramaswamy, "Prediction of Probable Genes by Fourier Analysis of Genomic Sequences", *Computer Applications in the Biosciences*, Vol. 13, pp 263 - 270, 1997.
- [5] D. Anastassiou, "Frequency-domain analysis of biomolecular sequences", *Bioinformatics*, Vol. 16, No. 12, pp. 1073-1081, 2000.
- [6] Paul Cristea, "Real and Complex Genomic Signals", *DSP*, pp 543-546, 2002.
- [7] B. Isaac, H. Singh, H. Kaur and G.P.S. Raghava, "Locating probable genes using Fourier Transform approach", *Bioinformatics*, Vol. 18, No. 1, 196-197, 2002.
- [8] W. Rudin, "Real and Complex Analysis", McGraw-Hill International Editions: Mathematics Series, 1987.
- [9] W.C. Liew, Yonghui Wu, H. Yan, Mengsu Yang, "Effective statistical features for coding and non-coding DNA sequence classification for yeast, C. elegans and human", *Int. J. Bioinformatics Research and Applications*, Vol. 1, pp 181 - 201, 2005.
- [10] Y. Wu, A.W.C. Liew, H. Yan and M. Yang, "Classification of short human exons and introns based on statistical features", *Physical Review E*, Vol. 67, pp 1-7, 2003.
- [11] M. Burset and R. Guigo, "Evaluation of gene structure prediction programs", *Genomic*, Vol. 34, pp 353-367, 1996.