GENE IDENTIFICATION USING THE $\mathcal{Z}$-CURVE REPRESENTATION

Ahmad Rushdi and Jamal Tuqan

Department of Electrical and Computer Engineering
University of California, Davis, CA 95616
aarushdi@ece.ucdavis.edu, tuqan@ece.ucdavis.edu

ABSTRACT

DSP based techniques have been recently proposed to identify the protein coding regions in a DNA strand by detecting the so-called period-3 component in the DNA spectrum. The DNA spectrum is computed after mapping the DNA symbolic sequence to numerical digital sequences. A typical choice of mapping is the Voss Representation. In this paper, we propose the use of a more elaborate mapping scheme, namely the $\mathcal{Z}$-curve representation. Using a multirate signal processing approach, we derive closed form expressions to compute the $\mathcal{Z}$-curve based DNA spectrum as well as mathematical conditions that characterize the coding regions. The derived formulas also prove that the $\mathcal{Z}$-curve representation yields essentially the same DNA spectrum as the one obtained using the Voss representation at a *lower computational cost*.

1. INTRODUCTION

A single DNA strand is represented as a sequence of letters that belong to the alphabet $\mathbb{F} = \{A,\ C,\ G,\ T\}$ denoting the four nucleotides (bases) in the DNA, *Adenine*, *Cytosine*, *Guanine*, and *Thymine*. Only segments of the DNA molecule contain relevant information for protein synthesis, namely the genes or the protein coding regions. A major goal of genomic research is to understand the nature of this information and its role in determining the particular gene function. This, in turn, requires the identification of the genes in a DNA sequence. By mapping the symbolic DNA sequence to a set of digital signals, Signal Processing techniques can be applied to identify these protein coding regions [1]. All of these approaches rely on detecting the existence of short range correlations in the nucleotide arrangement known as the "1/3 periodicity" or the "period-3" component in the **DNA spectrum** [1, 2, 3]. A number of researchers have indeed reported that the base sequence in the coding regions (exons in Eucaryotes) have a strong period-3 component [1, 2] while on the other hand no such peaks exist in non-coding regions (introns in Eucaryotes) [3, 4]. Given the period-3 behavior, it was concluded that the relative-height of the peak at $f = 1/3$ in the DNA spectrum is a good discriminator of coding potential and can therefore be used to distinguish between coding and non-coding regions. Although the period-3 has shown its merit in gene prediction for a number of genomes [1-4], a pending question is *whether the period-3 behavior of the DNA spectrum is not the result of the specific mapping scheme*, i.e., an artifact of the Voss representation itself rather than a biologically induced feature.

To address this question, we consider a more sophisticated mapping known as the **$\mathcal{Z}$-curve representation** [5, 6]. Using this specific map and by invoking multirate DSP principles, we derive closed form expressions for the DNA spectrum. The derived mathematical expressions are very simple to implement and generate a set of mathematical conditions that separate the coding regions from the non-coding ones. We also prove that the DNA spectrum based on the $\mathcal{Z}$-curve mapping is equivalent to the one obtained using the Voss representation. This suggests that the period-3 behavior is *independent* from the underlying mapping scheme. Finally, we show that the computational cost of the DNA spectrum using the $\mathcal{Z}$-curve mapping is much lower than in the Voss representation case. The above results imply in particular that it is *always more beneficial* to work with the $\mathcal{Z}$-curve map rather than the Voss representation when computing the DNA spectrum.

2. THE $\mathcal{Z}$-CURVE MAP

To generate the $\mathcal{Z}$-curve sequences, we first map a single DNA strand into four binary indicator sequences $x_\ell(n)$, $\forall\ \ell \in \{A,\ C,\ G,\ T\}$ where 1 indicates the presence of a base and 0 indicates its absence. For example, the mapping of the single DNA strand

5' ... A T G G T C T A A ... 3'

to the binary indicator sequence $x_A(n)$ is given by

$$x_A(n) = (..., 1, 0, 0, 0, 0, 0, 0, 1, 1, ...)$$

This simple and most popular mapping of a DNA sequence is known as the **Voss representation** [7]. From a biological perspective, the Voss representation characterizes the frequency of occurrence of each individual base ℓ in the DNA sequence. The sums of the samples for every $x_\ell(n)$ (as a function of n) are then computed to obtain

$$\ell(n) = \sum_{i=0}^{n} x_\ell(i), \quad \forall\ \ell \in\ \mathbb{F} \tag{1}$$

The four cumulative sequences $A(n)$, $C(n)$, $G(n)$, and $T(n)$ are in turn used to form the $\mathcal{Z}$-curve sequences $x(n)$, $y(n)$, and $z(n)$ as follows [6]

$$\begin{aligned} x(n) &= 2[A(n) - A(n-1) + G(n) - G(n-1)] - 1 \\ y(n) &= 2[A(n) - A(n-1) + C(n) - C(n-1)] - 1 \\ z(n) &= 2[A(n) - A(n-1) + T(n) - T(n-1)] - 1 \end{aligned}$$

for $n = 0, \ldots, N-1$ and $A(0) = G(0) = C(0) = T(0) = 0$. The above equations are however not suitable for our analysis and alternative expressions for $x(n)$, $y(n)$, and

$z(n)$ in terms of the Voss sequences are required. To derive these, recall that $A(n)$ is the summation of $x_A(n)$ from 0 to n and hence $A(n-1)$ is the summation of $x_A(n)$ from 0 to $n-1$. Then the cumulative difference $A(n)-A(n-1)$ is 1 if the n^{th} base of the sequence is an A and is 0 otherwise. It follows that

$$A(n) - A(n-1) = x_A(n) \tag{2}$$

Similar forms exist for C, G, and T. From equation (2), we can express $x(n)$, $y(n)$, and $z(n)$ as an **affine transformation** of the Voss sequences

$$\begin{bmatrix} x(n) \\ y(n) \\ z(n) \end{bmatrix} = 2 \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_A(n) \\ x_C(n) \\ x_G(n) \\ x_T(n) \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \tag{3}$$

We note that $x(n)$, $y(n)$, and $z(n)$ can only take the values of 1 or -1. For example, $x(n)$ has the value 1 *iff* the n^{th} sample of the sequence is either A or G and has the value -1 *iff* the n^{th} sample is either C or T. One advantage of the $\mathcal{Z}$-curve map that has partly motivated this study is that each of the three $\mathcal{Z}$-curve generated digital sequences *has a biological interpretation*. The first sequence $x(n)$ indicates the existence of either A or G which represents a differentiation between the purines/pyrimidines (**R/Y**) bases along the DNA strand. Similarly, the second sequence $y(n)$ represents the distribution of the amino/keto (**M/K**) types bases along the DNA sequence while the third sequence $z(n)$ represents the distribution of the strong/weak hydrogen bonds (**S/W**) [5]. A second advantage of the $\mathcal{Z}$-curve map is that $x(n)$, $y(n)$, and $z(n)$ are independent unlike the Voss representation where the four elementary sequences form a linearly dependent set since they add up to a sequence of ones. The removal of redundancy decreases the computational cost of the DNA spectrum as we show later.

3. THE VOSS DNA SPECTRUM

Assume that a DNA sequence $x(n)$ is of length N. The sliding window M-point DFT of $x(n)$ is defined by

$$X(m,k) \triangleq \sum_{n=0}^{M-1} x(n+m)e^{-j2\pi nk/M} \tag{4}$$

The starting point of the window is $m = 0, P, \ldots, (N-1)/P$ (we zero-pad $x(n)$ if $(N-1)/P \neq integer$) where P is the amount of window shift. If $P = 1$, then, the window slides one nucleotide at a time whereas if $P = 3$, the displacement of the window is on a codon basis (a codon is a triplet of nucleotides). To capture the period-3 component in $x(n)$, we let $M = 3L$ for some positive integer L and fix the frequency k to $L = M/3$. Equation (4) then becomes

$$X(m) \triangleq X(m,L) = \sum_{n=0}^{M-1} x(n+m)e^{-j2\pi n/3} \tag{5}$$

The polyphase representation for P = 3. Using the notation $x(m+n) \equiv x_m(n)$, we rewrite $X(m)$ as follows

$$\begin{aligned} X(m) &= \sum_{r=0}^{2} \sum_{n=r,r+3,\ldots}^{\lfloor \frac{N-1}{3} \rfloor} x_m(3n+r)e^{-j2\pi r/3} \\ &\triangleq \sum_{r=0}^{2} X_{m_r} e^{-j2\pi r/3} \end{aligned} \tag{6}$$

The signals $x_m(3n+r)$, $r \in \{0,1,2\}$, are termed respectively the first, second and third **polyphase components** of $x_m(n)$ [8] and can be generated by passing the signal $x_m(n)$ through the multirate blocking structure of Fig. 1. From Fig. 1 and to compute $X(m)$, we first generate the three polyphase components by shifting and downsampling with decimation ratio 3, sum their respective samples using the real FIR filter $H(z)$ (rectangular window), multiply the sums by the appropriate complex numbers and, finally add the three resulting quantities.

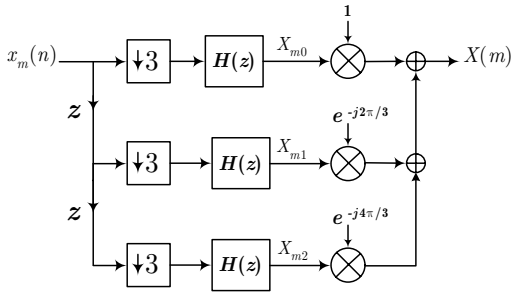

Figure 1: A multirate blocking structure with decimation ratio 3 and $H(z) = 1 + z^{-1} + z^{-2} + ... + z^{-(L-1)}$.

Fig. 2 shows the three polyphase components given by equation (6) as phasors. It can be seen that if $X_{m_0} = X_{m_1} = X_{m_2}$ as in Fig. 2(a), then the $m^{th}M/3$ DFT sample is zero, implying the existence of a non-coding region (intron). If this condition is not satisfied as in Fig. 2(b), a coding region (exon) is expected.

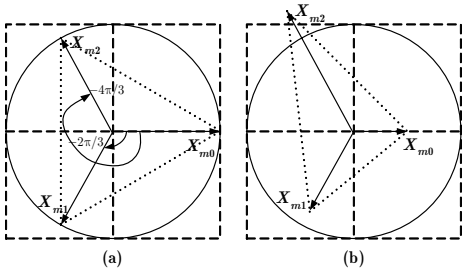

Figure 2: Phasor diagram of the polyphase components of $X(m)$ given by equation (6) (a) in case of equal amplitudes (b) in case of different amplitudes.

To compute the Voss DNA spectrum $S_v(m)$, we first find the polyphase components for each of the four Voss DFT sequences as given by equation (6). It follows that

$$X_{\ell m} \triangleq X_{\ell m_0} + X_{\ell m_1} e^{-j2\pi/3} + X_{\ell m_2} e^{-j4\pi/3} \quad \forall\, \ell \in \mathbb{F} \tag{7}$$

Given that the DNA spectrum is defined as

$$S_v(m) = |X_{Am}|^2 + |X_{Cm}|^2 + |X_{Gm}|^2 + |X_{Tm}|^2$$

and since $|X_{\ell m}|^2 = X_{\ell m} X^*_{\ell m}$ where * denotes complex conjugation, we can show using equation (7) that $|X_{\ell m}|^2 = 1/2 \sum_{r=0}^{2} [X_{\ell m_r} - X_{\ell m_q}]^2$. It follows that

$$S_v(m) = 1/2 \sum_{\ell \in \mathbb{F}} \sum_{r=0}^{2} [X_{\ell m_r} - X_{\ell m_q}]^2 \quad (8)$$

where $q = (r+1) \bmod 3$.

Computation Complexity. For a DNA strand of length N with the sliding window length M, the window starting point m ranges from 0 to $N-M$ and hence $N-M$ windows are to be tested for being a coding/non-coding region. For each window, $12 * 2 = 24$ multiplications are required. Therefore to test the whole sequence, the computation complexity is of order $\simeq 24\ N$ for relatively long sequences.

4. THE $\mathcal{Z}$-CURVE DNA SPECTRUM

To compute the $\mathcal{Z}$-curve DNA spectrum $S_z(m)$, we proceed as in section 3 by finding the polyphase representation of $x(n)$, $y(n)$, and $z(n)$ respectively. For example, for $x(n)$, we can write

$$\begin{aligned} X(m) &= \sum_{n=0}^{M-1} x(n) e^{-j2\pi n/3} \\ &= 2 \sum_{n=0}^{M-1} [x_A(n) + x_G(n)] e^{-j2\pi n/3} \\ &- \sum_{n=0}^{M-1} e^{-j2\pi n/3} \end{aligned} \quad (9)$$

Since $M = 3L$, the second summation in (9) is equal to 0. Using the polyphase representation, we find that

$$X(m) = 2 \sum_{r=0}^{2} (X_{Am_r} + X_{Gm_r}) e^{-j2\pi r/3} \quad (10)$$

Since $|X(m)|^2 = X(m) X^*(m)$ where * denotes complex conjugation, we get

$$|X(m)|^2 = 2 \sum_{r=0}^{2} [X_{Am_r} + X_{Gm_r} - X_{Am_q} - X_{Gm_q}]^2$$

where $q = (r+1) \bmod 3$. Similarly, we compute the spectral energies $|Y(m)|^2$ and $|Z(m)|^2$ of $y(n)$ and $z(n)$ respectively. The $\mathcal{Z}$-curve DNA spectrum is then given by

$$S_z(m) = 2 \sum_{\ell \in \grave{\mathbb{F}}} \sum_{r=0}^{2} [X_{Am_r} + X_{\ell m_r} - X_{Am_q} - X_{\ell m_q}]^2$$

where $q = (r+1) \bmod 3$ and $\grave{\mathbb{F}}$ is a subset of the nucleotides field $\mathbb{F}$, that is $\grave{\mathbb{F}} = \{C,\ G,\ T\} \subset \mathbb{F}$. Rearranging terms,

$$\begin{aligned} S_z(m) &= 4 S_v(m) \\ &+ 4 \sum_{r=0}^{2} (X_{Am_r} - X_{Am_q}) \sum_{\ell} (X_{Am_r} - X_{Am_q}) \end{aligned}$$

A simplification of the previous equation is obtained by observing that for $r \in \{0,1,2\}$, the quantity $[X_{A_r} + X_{G_r} + X_{C_r} + X_{T_r}]$ is equal to the total number of possible bases in the r^{th} codon position within the window. Since, in each codon within the window, the r^{th} position is always occupied by a nucleotide, the above quantity is a constant and is equal to 1/3 the window length. Using this conclusion, we can then derive that

$$S_z(m) = 4 S_v(m) \quad (11)$$

indicating that the $\mathcal{Z}$-curve DNA spectrum is simply a *scaled version* of the Voss DNA spectrum.

Conditions for the non-coding regions. We define the *combined binary indicator sequences* $X_{A\ell m_r} = X_{Am_r} + X_{\ell m_r}\ \forall\ r \in \{0,1,2\}$, and $\forall\ \ell \in \grave{\mathbb{F}}$. For example, X_{AGm_r} is 1 if the r^{th} codon position of the m^{th} window is either A or G and is 0 if the base at this position is either C or T. Using this definition, we can prove that

$$S_z(m) = 2 \sum_{\ell \in \grave{\mathbb{F}}} \sum_{r=0}^{2} (X_{A\ell m_r} - X_{A\ell m_q})^2 \quad (12)$$

for $q = (r+1) \bmod 3$. Note that, from equation (12), $S_z(m) \geq 0$. Moreover, $S_z(m) = 0$ if and only if every term composing $S_z(m)$ is equal to zero. This, in turn, produces the following conditions for the non-coding regions

$$\begin{aligned} &\Rightarrow \quad \frac{X_{A\ell m_r}}{X_{A\ell m_q}} = 1, \\ &\forall \quad r \in \{0,1,2\},\ q = (r+1)\ mod\ 3 \\ &and \quad \forall \quad \ell \in \grave{\mathbb{F}} \end{aligned} \quad (13)$$

Computation Complexity. For a DNA strand of length n with the sliding window length M, the window starting point m ranges from 0 to $N-M$ and hence $N-M$ windows are to be tested for being a coding/non-coding region. For each window, $9 * 2 = 18$ multiplications are required. Therefore to test the whole sequence, the computation complexity is of order $\simeq 18\ N$ for relatively long sequences.

5. SIMULATION RESULTS

To validate the mathematical derivations in the previous section, we compare the spectrums obtained from the $\mathcal{Z}$-curve and Voss representations of the DNA sequence of the gene F56F11.4 (Genbank accession number AF099922) in the C-elegans chromosome III. This gene has been used as a benchmark for different DSP gene detection schemes and is known to have five distinct exons. Applying the sliding window DFT method with a window length 351 on this gene using both representations resulted in *the same normalized spectra* shown in Fig. 3(a). To locate the protein coding regions, we use the threshold level $T = m_s + \gamma * \sigma_s$ where m_s is the mean of the DNA spectrum, σ_s is its standard deviation and, γ is an arbitrary real number. The threshold level T is therefore parameterized by γ. The value of γ is determined by using a learning paradigm over *known genes* and in our case, is found to be 0.6625. With γ fixed, binary decision plot $S_T(n)$ of the DNA spectrum of the F56F11.4 gene is shown in Fig. 3(b). In order to verify our results, *GenScan* and *Ensembl* have been used to

predict the exon-intron structure for the F56F11.4 gene. Their outputs are shown in Fig. 5 and Table. 1 respectively and match perfectly with our results. The above experiment demonstrates the potential of the method. An extensive study over a large number of genomes is outside the scope of this paper.

More on the $\mathcal{Z}$-curve map. Besides its computational advantage, the $\mathcal{Z}$-curve sequences can in general contribute towards the finding of important biological features. For example, for the gene F56F11.4, we found that the cumulative base sequences are dominated by $A(n)$ and $T(n)$ as shown in Fig. 4. It follows that $z(n)$, which is A and T dependent, is increasing at a much faster rate than $x(n)$ and $y(n)$. The DNA spectrum of $z(n)$ only would then represent the distribution of the bases of the strong/weak hydrogen bonds along the DNA sequence which can potentially help in locating areas in the sequence of increased density of cytosine-guanine, also known as *CpG islands*.

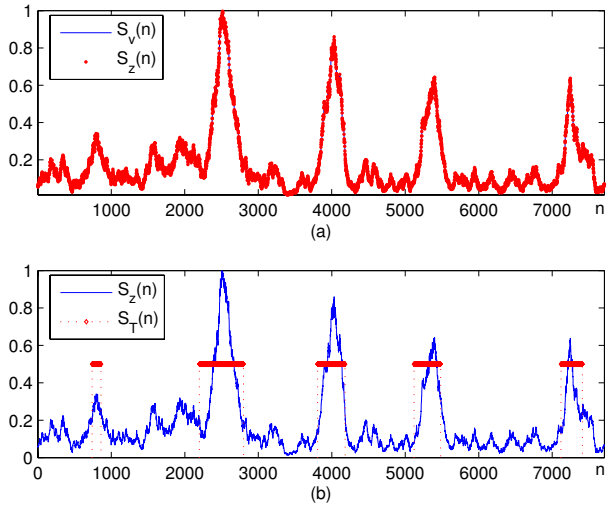

Figure 3: (a) Normalized DNA Spectrum of the gene F56F11.4 using the sliding window DFT method for both the Voss and $\mathcal{Z}$-curve representations (b) The $\mathcal{Z}$-curve DNA Spectrum of the gene F56F11.4 and the binary decision sequence output with $\gamma = 0.6625$.

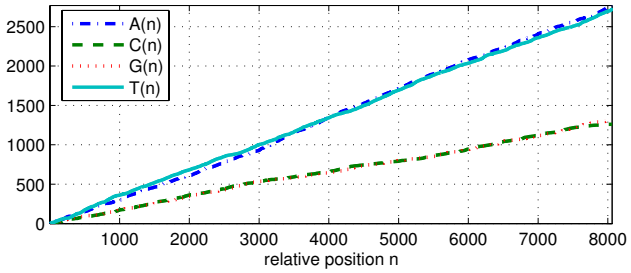

Figure 4: The Cumulative Base Sequences for the Gene F56F11.4 in the C-elegans chromosome III.

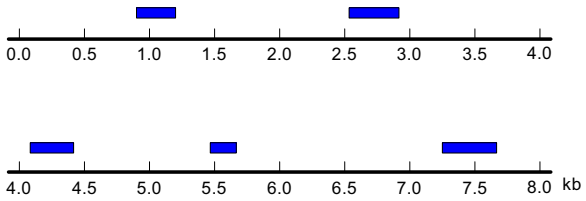

Figure 5: GenScan Coding Regions for F56F11.4

Region	Relative Position	Exon Length
F56F11.4a.1	929-1135	207
F56F11.4a.2	2528-2857	330
F56F11.4a.3	4114-4377	264
F56F11.4a.4	5465-5644	180
F56F11.4a.5	7255-7605	351

Table 1: Ensembl Coding Regions for F56F11.4

6. CONCLUDING REMARKS

A study of the $\mathcal{Z}$-curve map for the period-3 detection was presented. In specific, we have derived closed form expressions for the $\mathcal{Z}$-curve based DNA spectrum and mathematical conditions that characterize the coding regions from the non-coding ones. We have also proven that the $\mathcal{Z}$-curve DNA spectrum is essentially the same as the Voss based one. Finally, our analysis indicates that the computational cost involved in the evaluation of the $\mathcal{Z}$-curve spectrum is *less* than in the Voss case. This is particularly relevant given that some sequences can be several hundred thousand base pair long. One interesting related research question is whether all affine maps of the Voss sequences generate the same DNA spectrum.

7. REFERENCES

[1] D. Anastassiou, "Genomic signal processing," *IEEE Signal Proc. Mag.*, vol. 18, no. 4, pp. 8–20, July 2001.

[2] J. W. Fickett, "Recognition of protein coding regions in DNA sequences," *Nucleic Acids Research*, vol. 10, pp. 5303–5318, September 1982.

[3] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," *CABIOS*, vol. 13, pp. 263–270, June 1997.

[4] B. D. Silverman and R. Linsker, "A measure of DNA periodicity," *Journal of Theoretical Biology*, vol. 118, no. 3, pp. 295–300, February 1986.

[5] R. Zhang and C.T. Zhang, "Z curves, an intuitive tool for visualizing and analyzing the DNA sequences," *J. on Biom. Struc. Dyn.*, vol. 11, pp. 767–782, July 1994.

[6] M. Yan, Z. S. Lin, and C. T. Zhang, "A new Fourier Transform approach for protein coding measure based on the format of the Z-curve," *Bioinformatics*, vol. 14, no. 8, July 1998.

[7] R. F. Voss, "Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences," *Phy. Rev. Lett.*, vol. 68, no. 25, pp. 3805–3808, June 1992.

[8] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, 1993.