A DSP PERSPECTIVE TO THE PERIOD-3 DETECTION PROBLEM

Jamal Tuqan and Ahmad Rushdi

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

ABSTRACT

Many Signal Processing techniques have been introduced in the past to identify the protein coding regions by detecting the so-called period-3 component in the DNA spectrum. However, a solid understanding of this observed phenomenon and its underlying mechanism from a DSP perspective has been missing from the literature. We therefore propose a novel DSP model that i) clearly explains the intricate operation of the DNA spectrum, ii) allows the derivation of new DNA spectrum expressions which, in turn, generalize and unify previous work and iii) suggests an efficient way to improve the detection of protein coding regions by computing a *filtered* spectrum.

1. INTRODUCTION

A central problem of genomic research is to find the number and locations of the genes (exons in eucaryotes) and, their exact boundaries (start & stop codons and splice sites). Special features are needed to discriminate between the protein coding regions (genes) and the non-coding ones. In this work, we study in depth one such feature, namely the period-3 component of the DNA spectrum.

A single DNA strand is a sequence of nucleotides (bases) where each base belongs to the alphabet $\mathbb{F} = \{A,\ C,\ G,\ T\}$. To perform signal Processing operations on the DNA sequence, numerical values are assigned to each character in $\mathbb{F}$. A typical assignment is the so-called Voss representation where four binary sequences, $x_l(n)$, $l \in \mathbb{F}$, are generated with 1 indicating the existence of the base l at position n [1]. Assume that a DNA sequence has length N. The sliding window M-point DFT for each $x_l(n)$ is

$$X_l(n,k) \triangleq \sum_{m=0}^{M-1} x_l(n+m)e^{-j2\pi mk/M} \qquad (1)$$

where the starting point of the window $n = 0,\ 1, \ldots, N-1$ and, $M = 3L$ where L is a positive integer. The frequency spectrum of the given DNA sequence is therefore given by

$$S(n,k) = \sum_{l\in\mathbb{F}} |X_l(n,k)|^2$$

It was observed by many researchers that the spectrum of protein coding regions has typically a peak at $k = L$ whereas no significant peaks appear in the spectrum of non-coding regions [2]. The **DNA Spectrum**, $S(n)$, can therefore be used to distinguish the coding regions from the non-coding ones and is defined by

$$S(n) = \sum_{l\in\mathbb{F}} |X_l(n,L)|^2 = \sum_{l\in\mathbb{F}} |\sum_{m=0}^{N-1} x_l(n+m)e^{-j2\pi m/3}|^2 \qquad (2)$$

where $n = 0,\ P,\ \ldots,\ (N-1)/P$ (we zero-pad $x_l(n)$ if $(N-1)/P \neq integer$) and P is the amount of window shift. From (2), it is easy to show that the sequence $X_l(n)$ $\triangleq X_l(n,L)$ is obtained by passing $x_l(n)$ through the filter

$$f(n) = \begin{cases} e^{j2\pi n/3} & 0 \leq n \leq M-1 \\ 0 & otherwise \end{cases}$$

as shown below. The box labelled $\downarrow P$ is a downsampler with decimation ratio P [3], and $F(z) = \mathcal{Z}\{f(n)\}$.

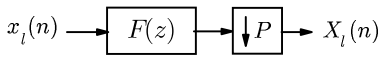

Figure 1: Digital Filtering Model for Period-3 Detection

2. THE DSP MODEL

We first observe that $F(z)$ in Figure 1 can be expressed as $C(z)H(z^3)$ where $C(z) = 1 + e^{j2\pi/3}z^{-1} + e^{j4\pi/3}z^{-1}$ and $H(z^3) = 1 + z^{-3} + ... + z^{-(M-1)}$. The elegance of this two-stage filter model is that *it clarifies the inner workings* of the sliding window DFT. To see this, note that the filter $H(z)$ is the standard rectangular window and has a low pass frequency response with a -13 db attenuation. Its *interpolated* version, $H(z^3)$, produces frequency images at $\omega = 0,\ 2\pi/3$ and $4\pi/3$. The complex filter $C(z)$ has zeros at $\omega = 0$ and $\omega = 4\pi/3$ and thus attenuates the images at these frequencies. The resulting filter $F(z)$ is, as expected, a complex band pass filter with center frequency at $2\pi/3$. With the two-stage filter model in mind, we propose the structure of Figure 2. The signals $x_{l0}(n)$, $x_{l1}(n)$ and, $x_{l2}(n)$ are termed the first, second, and third polyphase components of $x_l(n)$ respectively [3]. From a biological perspective, $x_{lr}(n)$ is the indicator sequence of base l in the r^{th} codon position $(r = 0, 1, 2)$. When $H_r(z)$ are the rectangular window $H(z) = 1 + z^{-1} + \ldots + z^{-(L-1)}$ for $r = 0, 1, 2$, we can show that $X_l(n)$ is the DNA spectrum of section 1 with $P = 3$. In the more general case, we have

$$X_l(n) = X_{l0}(n) + e^{j2\pi/3}X_{l1}(n) + e^{j4\pi/3}X_{l2}(n) \qquad (3)$$

where *the filtered polyphase components* $X_{lr}(n) = x_{lr}(n) * h_r(n)$ for $r = 0, 1, 2$ and $*$ denotes linear convolution.

1-4244-0385-5/06/$20.00 ©2006 IEEE

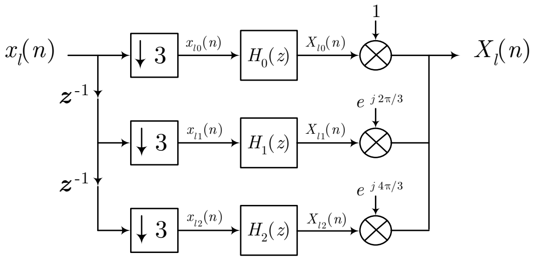

Figure 2: The new DSP model

3. MAIN RESULTS OF THE PAPER

We assume here the general FIR case where the polyphase filters $H_r(z) = \alpha_{r0} + \alpha_{r1} z^{-1} + \ldots + \alpha_{r(L-1)}\ z^{-(L-1)}$. **Closed form expression for the spectrum.** From (3), we can prove that

$$S(n) = \frac{1}{2} \sum_{l \in \mathbb{F}} \sum_{r=0}^{2} [X_{lr}(n) - X_{lq}(n)]^2 \qquad (4)$$

where $q = (r+1) \bmod 3$. The above equation shows in specific that the DNA spectrum is *completely characterized* by the filtered polyphase components and can be computed directly from these sequences. The formula also indicates that the value of the spectrum is independent from the particular reading frame. Finally, when $H_r(z) = H(z) = 1 + z^{-1} + \ldots + z^{-(L-1)}$ for $r = 0, 1, 2$, equation (4) reduces to a recent result derived in [4] using a parsing approach. **Biological Meaning of the DNA spectrum.** By algebraically manipulating (4), we can show that

$$S(n) = \frac{3}{2} \sum_{l \in \mathbb{F}} \sum_{r=0}^{2} [X_{lr}(n) - \bar{X}_l(n)]^2 \qquad (5)$$

where $\bar{X}_l(n) = \frac{1}{3}[X_{l0}(n) + X_{l1}(n) + X_{l2}(n)]$. In the *special case* of the rectangular window, equation (5) is equivalent to the *position asymmetry* measure [5]. This latter measure has a very nice biological interpretation. In specific, by observing that $X_{lr}(n)$ is the number of occurrences of base l in the r^{th} $(r = 0, 1, 2)$ position of the codon at window location n, the term $X_{lr}(n) - \bar{X}_l(n)$ represents the relative abundance of base l in codon position r with respect to the average frequency of occurrence of base l in all three codon positions. The more uneven the relative abundances are, the more likely the processed section of the DNA sequence is a protein coding region. In the general case, the r^{th} codon in the window is weighted by α_{rv}. **The polyphase Filtered DNA spectrum.** Unlike previous work where a smoothed DNA spectrum is obtained by changing $F(z)$ [4, 6], we propose instead to use $H_r(z)$ to filter the spectrum. This new approach provides a number of advantages over the former one. First, when $F(z)$ is modified, equations (4) and (5) are not valid anymore and the spectrum looses its biological significance. By contrast, the expressions in (4) and (5) are independent from the form of $H_r(z)$. Second, designing the complex band-pass filter $F(z)$ is more elaborate than designing a real low pass filter $H_r(z)$. Furthermore, the computation of $X_l(n)$ using $F(z)$ requires complex arithmetic whereas that of $S(n)$ using (4) or (5) involves *only* real operations. Finally, changing $H_r(z)$ generates a filtered DNA spectrum that not only smoothes the standard (rectangular window based) one but also identifies potential exon regions completely missing in the former case. This is clearly illustrated in Figure 3 where the standard DNA spectrum and the filtered one using a Blackman window with $M = 702$ are depicted for the zeta globin gene (ECZGL2) and the muscle actin gene (HROMA4A) acquired from the Burset-Guigo database. To find the exons, potential coding regions are first determined by applying a threshold $T = \mu_s + 0.5\sigma_s$ to $S(n)$ where μ_s and σ_s are the mean and standard deviation of $S(n)$ respectively. Once these regions are obtained, the boundaries of the initial, internal and terminal exons are found by searching locally for the consensus nucleotides. The discovered exons are outlined in Figures 3(b) and (d) by the (green) dashed binary curve, and with the exception of the internal exon in the zeta globin gene, match perfectly with the true exons represented in Figure 3 by the (black) solid binary curve. By contrast, the performance of the standard DNA spectrum was poor, producing many false positives and negatives.

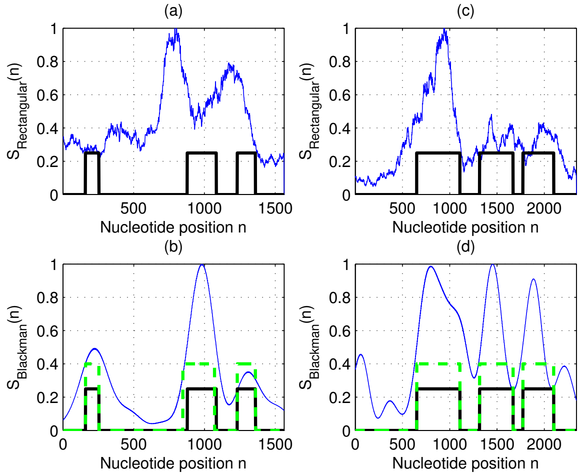

Figure 3: (a)-(b) and (c)-(d) Rectangular and Blackman window based spectrums for the zeta globin and the muscle actin genes respectively

4. REFERENCES

[1] R. F. Voss, "Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences," *Physics Reviews Letters*, vol. 68, pp. 3805–3808, June 1992.

[2] J. W. Fickett, "Recognition of protein coding regions in DNA sequences," *Nuc. Aci. Res.*, vol. 10, Sep. 1982.

[3] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Englewood Cliffs, NJ: Prentice Hall, 1993.

[4] S. Datta and A. Asif, "A fast DFT based gene prediction algorithm for identification of protein coding regions," in *Proc. of the ICASSP*, 2005, pp. 113–116.

[5] J. W. Fickett and C. S. Tung, "Assessment of protein coding measures," *Nuc. Aci. Res.*, vol. 20, Sep. 1992.

[6] P. P. Vaidyanathan and B.-J. Yoon, "Gene and exon prediction using all pass-based filters," *Gensips*, 2002.