GENE AND EXON PREDICTION USING ALLPASS-BASED FILTERS

P. P. Vaidyanathan and Byung-Jun Yoon
Dept. of Electrical Engineering, California Institute of Technology
Pasadena, CA 91125, USA
ppvnath@systems.caltech.edu bjyoon@caltech.edu

ABSTRACT

It is well-known that the protein-coding regions of DNA sequences exhibit a period-3 behavior due to codon structure. These regions are often identified with the help of techniques such as the windowed DFT. It has been demonstrated in the past that identification of the period-3 regions helps in predicting the gene locations, and in fact allows the prediction of specific exons within the genes of eucaryotic cells. In this paper we introduce a simple and efficient scheme for identifying the period-3 regions of DNA sequences based on antinotch IIR filters instead of the DFT. These filters can be implemented very efficiently using the one-multiplier Gray and Markel lattice structure.[1]

1. INTRODUCTION

It is well-known that base sequences in the protein-coding regions of DNA molecules exhibit a period-3 pattern because of the codon structure involved in the translation of base sequences into amino acids [9], [10]. For eucaryotes (cells with nucleus) this periodicity has mostly been observed within the exons (coding subregions inside the genes) and not within the introns (noncoding subregions in the genes). There are theories explaining the reason for such periodicity, but there are also exceptions to the phenomenon. For example, certain rare genes in S. cerevisiae (also called baker's yeast) do not exhibit this periodicity [9]. Furthermore for procaryotes (cells without a nucleus), and some viral and mitochondrial base sequences, such periodicity has even been observed in noncoding regions [5]. For this and many other reasons, gene prediction is a very complicated problem (see the review article by Fickett [3]). Nevertheless, many researchers have regarded the period-3 property to be a good (preliminary) indicator of gene location. Techniques which exploit this property for gene prediction proceed by computing the discrete Fourier transform (DFT), which is expected to exhibit a peak at the frequency $2\pi/3$ due to the periodicity. In fact this techique has also been used to isolate exons within the genes of eucaryotic cells [2, 9]. The periodic behavior indicates strong short-term correlation in the coding regions, in addition to the long-range correlation or $1/f$-like behavior exhibited by DNA sequences in general [5,7,11].

In this paper we provide an efficient mechanism for the identification of regions exhibiting period-3 behavior. This is based on digital IIR filtering. Specifically, the output of an antinotch filter, with a sharp gain at the frequency $2\pi/3$ provides this information as a function of base location. This is more efficient than the computation of the DFT based on overlapping windows. In a way it can be regarded as a recursive implementation of an FIR filter. The antinotch filter can be implemented very efficiently with the help of an allpass filter, which in turn can be realized with a lattice structure having only two multipliers. One of these multipliers controls the sharpness of the filter peak, by controlling the pole radius. The antinotch frequency is fixed at $2\pi/3$ with the help of the other multiplier. There is a compromise between the sharpness of the notch filter and the base-domain resolution achievable, but the method appears to be promising. The performance of the scheme will be demonsrated on gene sequences taken from C.elegans in the public geneomic data base.

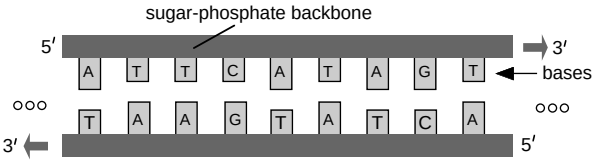

Figure 1. The DNA double helix (linearized schematic).

2. BASE-PERIODICITY IN CODING REGIONS

An overview of some of the important aspects of the DNA molecule from the signal processing view point can be found in the article by Anastassiou [2]. Figure 1 demonstrates a simple schematic for part of a DNA molecule [1], with the double helix straightened out for simplicity. The four bases or nucleotides attached to the sugar phosphate backbone are denoted with the usual letters A, C, G, and T (respectively, adenine, cytosine, guanine, or thymine). Note that the base A always pairs with T, and C pairs with G. The two strands of the DNA molecule are therefore complementary to each other. The forward genome sequence correspond to the upper strand of the DNA molecule, and in the example shown this is $ATTCATAGT$. Note that the ordering is from the so-called $5'$ to the $3'$ end (left to right). The complementary sequence corresponds to the bottom strand, again read from $5'$ to $3'$ (right to left).

[1]Work supported in part by the ONR grant N00014-99-1-1002, USA.

This is $ACTATGAAT$ in our example. DNA sequences are always listed from the $5'$ to the $3'$ end because, they are scanned in that direction when triplets of bases (codons) are used to signal the generation of amino acids. Typically, in any given region of the DNA molecule, at most one of the two strands is active in protein synthesis (multiple coding areas, where both strands are separately active, are rare).

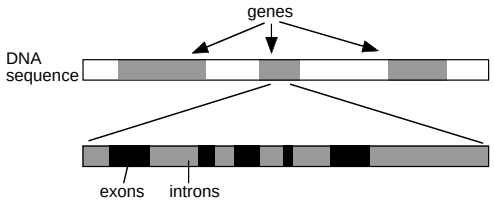

Figure 2. A DNA sequence has genes and intergenic regions. The genes of eucaryotes have exons (protein coding regions) and introns.

Figure 2 shows various regions of interest in a DNA sequence, which can be divided into genes and intergenic spaces. The genes are responsible for protein synthesis. Even though all the cells in an organism have identical genes, only a selected subset is active in any particular family of cells. A gene, which for our purposes is a sequence made up from the four bases, can be divided into two subregions called the exons and introns. (Procaryotes, which are cells without a nucleus, do not have introns). Only the exons are involved in protein-coding. The bases in the exon region can be imagined to be divided into groups of three adjacent bases. Each triplet is called a codon. Evidently there are 64 possible codons. Scanning the gene from left to right, a codon sequence can be defined by concatenation of the codons in all the exons. Each codon (except the so-called stop codon) instructs the cell machinery to sythesize an amino acid. The codon sequence therefore uniquely identifies an amino acid sequence which defines a protein. Since there are 64 possible codons but only 20 amino acids, the mapping from codons to amino acids is many-to-one. The introns do not participate in the protein synthesis because they are removed in the process of forming the RNA molecules which carry the genetic code to the protein machinery outside the nucleus.

It has been observed more than two decades ago that the base sequence in the coding regions (exons) have a strong period-3 component (an observation perhaps attributable to Trifonov and Sussman [10]). Some authors have claimed that this is due to nonuniform codon usage: even though there are several codons which could code a given amino acid, they are not used with uniform probability, and this creates a codon bias. There is an excess of guanine (G) in position 1, leading to strong period 3 oscillation [4]. The work by Tiwari, et al. [1997] seems to indicate that this explanation is not complete. Indeed, these authors "synthesize genes" by starting from proteins and mapping aminoacids back to codons. In this reverse mapping process, they assign "uniform probability" to the different codons that might lead to a given amino acid. The resulting pseudo gene, by construction, is free from introns (like cDNA [1]), and it has been found that the period 3 property is still in tact! Tiwari, et al. also observe that some genes do not exhibit period 3 at all in S. cerevisiae (e.g., genes of the mating type locus).

3. SPECTRUM OF BASE SEQUENCES

Many researchers have regarded the period-3 property to be a good (preliminary) indicator of gene location, in fact exon location. To perform gene prediction based on this, one defines indicator sequences for the four bases and computes the DFTs of short segments of these. Given a DNA sequence, the *indicator sequence* for the base A is a binary sequence of the form $x_A(n) = 000110111000101010\ldots$ where 1 indicates the presence of an A and 0 indicates its absence. The indicator sequences for the other bases are defined similarly. It is clear that the sequence $111111\ldots$ is obtained by adding the four indicator sequences. The DFT of a length-N block of $x_A(n)$ is defined as

$$X_A[k] = \sum_{n=0}^{N-1} x_A(n) e^{-j2\pi kn/N}, \quad 0 \le k \le N-1,$$

where we have assigned number 0 to the beginning of the block. The DFTs $X_T[k], X_C[k]$, and $X_G[k]$ are defined similarly. The period-3 property of a DNA sequence implies that the DFT coefficients corresponding to $k = N/3$ are large. Thus if we take N to be a multiple of 3 and plot

$$S[k] \triangleq |X_A[k]|^2 + |X_T[k]|^2 + |X_C[k]|^2 + |X_G[k]|^2 \quad (1)$$

then we should see a peak at the sample value $k = N/3$ as demonstrated in many papers (e.g., [9]). While this is generally true, the strength of the peak depends markedly on the gene. It is sometimes very pronounced, sometimes quite weak.

Notice that a calculation of the DFT at the single point $k = N/3$ is sufficient. The window can then be slided by one or more bases and $S[N/3]$ recalculated. Thus, we get a picture of how $S[N/3]$ evolves along the length of the DNA sequence. It is necessary that the window length N be sufficiently large (typical window sizes are a few hundreds, eg., 351, to a few thousands) so that the periodicity effect dominates the background $1/f$ spectrum which makes its strong presence in DNA sequences [7], [11]. However a long window implies longer computation time, and also compromises the base-domain resolution in predicting the exon location.

4. THE IIR ANTINOTCH DNA-FILTER

Let $H(z)$ be a digital filter with magnitude response $|H(e^{j\omega})|$ as demonstrated in Fig. 3. The response has a sharp peak at $\omega = 2\pi/3$. If the indicator sequence $x_A(n)$ is passed through such a filter, then in coding regions we expect the output to be large because of the period-3 property described in Sec. 2. Such a filter can therefore be used to predict the coding regions. The duration for which the impulse response $h(n)$ is significant is analogous to the window length N in the DFT computation of Sec. 3, and determines the tradeoff between base-domain resolution and amplification factor for the period-3 property.

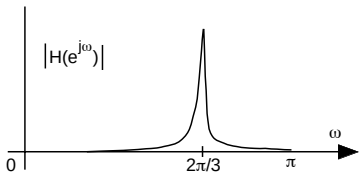

Figure 3. An antinotch filter.

FIR filters with such responses would require long impulse responses, implying more computations. IIR filters require much shorter orders and can be very efficient here. Such filters can be built from second order allpass filters. A second order real coefficient allpass filter with poles at $Re^{\pm j\theta}$ has transfer function

$$A(z) = \frac{R^2 - 2R\cos\theta z^{-1} + z^{-2}}{1 - 2R\cos\theta z^{-1} + R^2 z^{-2}} \tag{2}$$

where $R^2 < 1$ for stability. Notice that the numerator is the mirror image of the denominator so that the zeros are at the reciprocal locations $1/Re^{\pm j\theta}$. With $D(z)$ denoting the denominator, $A(e^{j\omega}) = e^{-2j\omega}D^*(e^{j\omega})/D(e^{j\omega})$ which proves the allpass property $|A(e^{j\omega})| = 1$. Next consider a filter of the form

$$G(z) = \frac{1 + A(z)}{2} \tag{3}$$

This can be simplified to the form

$$G(z) = K\Big(\frac{1 - 2\cos\phi z^{-1} + z^{-2}}{1 - 2R\cos\theta z^{-1} + R^2 z^{-2}}\Big)$$

where $K = (1 + R^2)/2$ and $\cos\phi = 2R\cos\theta/(1 + R^2)$. This shows that the zeros of $G(z)$ are on the unit circle at the angle ϕ. As the poles move closer to the unit circle (i.e., for $R \to 1$), then $\phi \approx \theta$ and the filter $G(z)$ has poles and zeros close together. Their effect therefore cancels for frequencies sufficiently away from ϕ, and the magnitude response of $G(z)$ is as demonstrated in Fig. 5: it approximates unity everywhere except in the neighbourhood of the notch frequency ϕ. Such a filter is called a notch filter. We now show that if we define a new filter as the difference

$$H(z) = \frac{1 - A(z)}{2} \tag{4}$$

then the response $|H(e^{j\omega})|$ has the antinotch property. For this observe that the filters $G(z)$ and $H(z)$ can together be expressed as

$$\begin{bmatrix} G(z) \\ H(z) \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 \\ A(z) \end{bmatrix} \tag{5}$$

Using the fact that the 2×2 matrix is unitary (upto scale) and that $|A(e^{j\omega})| = 1$, it follows immediately that

$$|G(e^{j\omega})|^2 + |H(e^{j\omega})|^2 = 1. \tag{6}$$

That is, the filters $G(z)$ and $H(z)$ are power complementary. Since $G(z)$ has the notch behavior, the antinotch property of $H(z)$ follows from this.

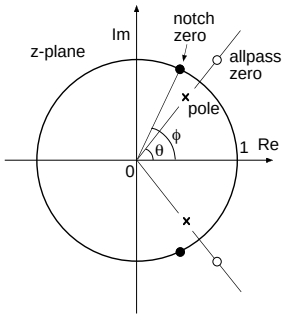

Figure 4. Poles and zeros of the notch filter $G(z)$ and the allpass filter $A(z)$

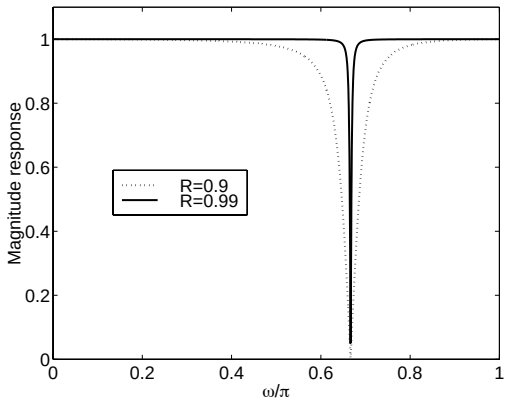

Figure 5. Notch filter responses for two values of R.

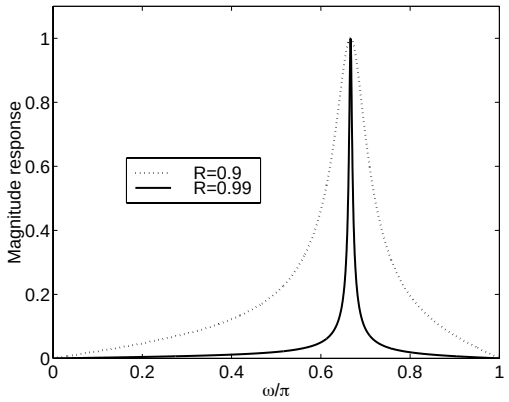

Figure 6. Antinotch filter responses for two values of R.

Figure 5 shows examples of the notch response $|G(e^{j\omega})|$ for two pole radii R, and notch frequency $2\pi/3$. The power complementary antinotch response $|H(e^{j\omega})|$ is shown in Fig. 6. We can make the response arbitrarily sharp by making R close to unity. However, the effects of roundoff

noise will eventually become noticeable if R is too close to unity, and moreover the significant part of the impulse response $h(n)$ will become very long, compromising the base-domain resolution in the prediction of gene locations.

The allpass filter $A(z)$ can be implemented with either the direct form structure or the cascaded lattice structure [6], [12]. The lattice structure with one-multiplier sections [12] is especially attractive [8], and Fig. 7 shows the implementation of $H(z)$ using this lattice. The multipliers in this structure are R^2 and $-\cos\phi$. Since the antinotch frequency is $\phi = 2\pi/3$ we have $-\cos\phi = 2^{-1}$. So the only significant multiplier is R^2, and controls the antinotch quality without affecting the frequency ϕ (Fig. 6). Thus we can adaptively and readily adjust R^2 depending on the base-domain resolution desired.

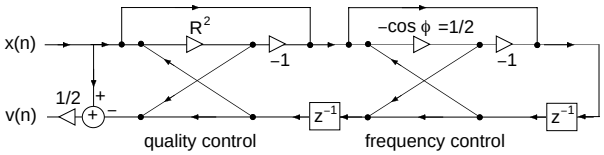

Figure 7. Lattice structure for implementing the antinotch filter $H(z) = V(z)/X(z)$.

5. EXAMPLES AND CONCLUSIONS

With the indicator sequence $x_A(n)$ taken as input, let $y_A(n)$ denote the output of the antinotch filter $H(z)$. With similar notation for the other bases, define

$$Y[n] = |y_A(n)|^2 + |y_T(n)|^2 + |y_C(n)|^2 + |y_G(n)|^2$$

Note that n should be interpreted as base location. $Y[n]$ is analogous to the traditional DNA spectrum $S[k]$ evaluated at $k = N/3$. A plot of $S[N/3]$ as a function of base location is shown in Fig. 8 (top) for the gene F56F11.4 in C-elegans (base number 7021 — 15080 in chromosome III; accession number AF099922). This gene has five exons, and the last four of them show clear peaks in the plot. The peak due to the first exon is unfortunately not dominant. The quantity $Y[n]$ computed for the same gene (pole radius $R = 0.992$) is also shown in the figure (bottom). The first exon is also visible now, in the sense that it dominates spurious peaks. Thus the allpass-antinotch appears to work very well, and furthermore offers some advantages in implemenation. A more detailed study for several genes from different organisms will be reported in future.

6. REFERENCES

[1] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential cell biology*, Garland Publishing Inc., New York, 1998.

[2] D. Anastassiou, "Genomic signal processing," IEEE Signal Processing Magazine, pp. 8–20, July 2001.

[3] J. W. Fickett, "The gene prediction problem: an overview for developers", Computers Chem., vol. 20, no. 1, pp. 103–118. 1996.

[4] H. Herzel, E. N. Trifonov, O. Weiss, and I. Groβe, "Interpreting correlations in biosequences," Physica A, vol. 249, pp. 449–459, 1998.

[5] W. Li, "The study of correlation structures of DNA sequences: a critical review", Computers Chem., vol. 21, no. 4, pp. 257–271, 1997.

[6] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing,* Prentice Hall, Inc., NJ, 1999.

[7] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, "Long-range correlations in nucleotide sequences," Nature, vol. 356, pp. 168–170, March 1992.

[8] P. A. Regalia, S. K. Mitra, and P. P. Vaidyanathan, "The digital allpass filter: a versatile signal processing building block," Proc. IEEE, pp. 19–37, Jan. 1988.

[9] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," CABIOS, vol. 13, no. 3, pp. 263–270, 1997.

[10] E. N. Trifonov, and J. L. Sussman, "The pitch of chromatin DNA is reflected in its nucleotide sequence", Proc. of the Nat. Acad. Sci., USA, vol. 77, pp. 3816–3820, 1980.

[11] R. F. Voss, "Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences," Physical Review Letters, vol. 68, no. 25, pp. 3805–3808, June 1992.

[12] P. P. Vaidyanathan, *Multirate systems and filter banks,* Prentice Hall, Inc., 1993.

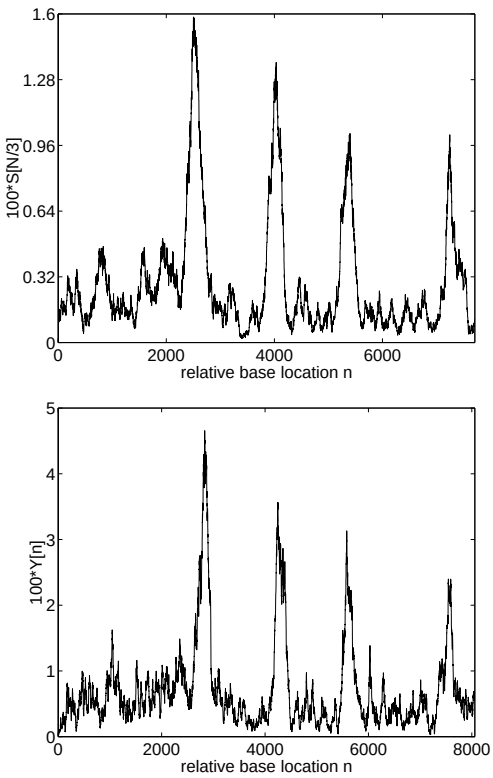

Figure 8. Top plot: the DFT based spectrum $S[N/3]$ for gene F56F11.4 in the C-elegans chromosome III. Bottom plot: the antinotch filter output for the same gene.