Proc. IEEE Asilomar Conf. Signals, Systems, and Computers, Nov. 2002, Monterey, CA

Digital filters for gene prediction applications

P. P. Vaidyanathan and Byung-Jun Yoon

Dept. Electrical Engr., Caltech. Pasadena, CA
ppvnath@systems.caltech.edu bjyoon@caltech.edu

Abstract. It has been observed by many researchers that the protein-coding regions of DNA sequences exhibit a period-3 behavior due to codon structure. 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. Traditionally these regions are identified with the help of techniques such as the windowed DFT. In this paper we consider the use of efficient digital filters for the same purpose. The filters can be designed not only to extract the period-3 component, but at the same time effectively eliminate the background $1/f$ spectrum exhibited by nearly all DNA sequences.[1]

I. 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 [11], [12]. For eucaryotes (cells with nucleus) this periodicity has mostly been observed within the exons (coding subregions inside the genes [1]) 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. 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) with a sliding window. This is expected to exhibit a peak at the frequency $2\pi/3$ due to the periodicity. This technique has successfully been used to identify exons within the genes of eucaryotic cells [2, 11]. 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 of many organism in general [6,9,15].

Digital signal processing techniques offer more efficient ways to identify regions of the DNA exhibiting periodic behavior. Such methods have typically not been used in the biotechnology community. For example, digital bandpass filters with a narrow passband are often very effective in extracting the period-3 information and attenuating the $1/f$ behavior. In this paper we describe a number of methods to obtain such filters, and examine their performance on DNA sequences from the genome database.

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

II. PERIODICITY IN CODING REGIONS

Figure 1(a) demonstrates a simple schematic for part of a DNA molecule [1], with the double helix straightened out for convenience. The four bases or nucleotides attached to the sugar phosphate backbone are denoted with the usual letters A, C, G, and T. The forward genome sequence $\ldots ATTCATAGT \ldots$ corresponds to the upper strand of the DNA molecule. Note that the ordering is from the so-called $5'$ to the $3'$ end (left to right).

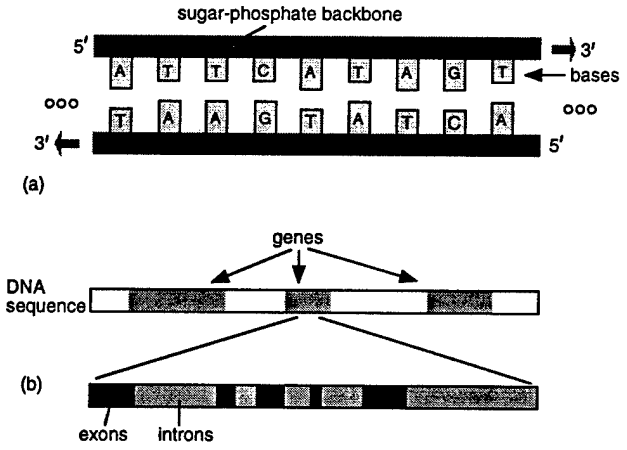

Figure 1. (a) The DNA double helix (linearized schematic), and (b) various regions in a DNA molecule.

As shown in Figure 1(b), a DNA sequence can be divided into genes and intergenic spaces. The genes are responsible for protein synthesis. A gene 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. 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 synthesize 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.

It has been observed more than two decades ago [12] that the base sequence in the coding regions (exons) have a strong period-3 component. 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 [5]. The work by Tiwari, et al. [11] 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 has been found to retain the period 3 property!

III. DNA SPECTRUM VERSUS DNA FILTERING

To perform gene prediction based on the period-3 property, 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, e.g.,

$$x_A(n) = 000110111000101010\ldots$$

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 the number $n = 0$ 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., [11]). 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 [9], [15]. However a long window implies longer computation time, and also compromises the base-domain resolution in predicting the exon location.

Digital filtering method. The sliding window method can be regarded as digital filtering followed by a decimator which depends on the separation between adjacent positions of the window [3, 13]. The filter itself has a very simple impulse response

$$w(n) = \begin{cases} e^{j\omega_0 n} & 0 \le n \le N-1 \\ 0 & \text{otherwise.} \end{cases}$$

This is a bandpass filter with passband centered at $\omega_0 = 2\pi/3$ and minimum stopband attenuation of about 13 dB (Fig. 2). This tells us that if we pay more careful attention to the design of the digital filter, we can isolate the period-3 behavior from background information such as $1/f$ noise more effectively. We can also use efficient methods to design and implement the filter, thereby reducing computational complexity.

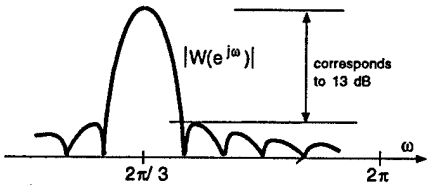

Figure 2. The filtering effect of DFT computation.

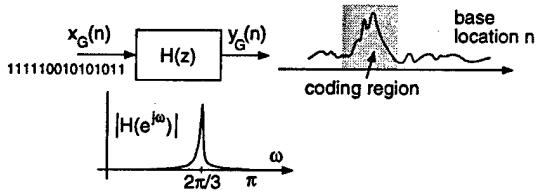

Figure 3. A digital filter $H(z)$ with indicator sequence $x_G(n)$ as its input.

Consider a narrow band bandpass digital filter $H(z)$ with passband centered at $\omega_0 = 2\pi/3$. With the indicator sequence $x_G(n)$ taken as input, let $y_G(n)$ denote its output. Note that n should be interpreted as base location. In the coding regions, the sequence $x_G(n)$ is expected to have a period-3 component, which means that it has large energy in the filter passband. So we expect the output $y_G(n)$ to be relatively large in the coding regions as demonstrated in Fig. 3. With similar notation for the other bases, define

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

A plot of this function can be used as a preliminary indicator of coding regions. The narrow band filter $H(z)$ can be regarded as an **antinotch filter** (i.e., complement of a notch). We now describe some efficient ways to design and implement such filters.

IV. IIR ANTINOTCH FILTERS

The use of IIR antinotch filters for gene prediction was proposed in [14]. Such IIR filters can be obtained by starting

from a second order allpass filter

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

which has poles at $Re^{\pm j\theta}$ and zeros at $1/Re^{\pm j\theta}$. Thus, consider a filter bank with two filters $G(z)$ and $H(z)$ defined according to

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

Then $G(z)$ has the form

$$G(z) = K\left(\frac{1 - 2\cos\omega_0 z^{-1} + z^{-2}}{1 - 2R\cos\theta z^{-1} + R^2 z^{-2}}\right)$$

where

$$\cos\omega_0 = \frac{2R\cos\theta}{1 + R^2}$$

This shows that $G(z)$ is a **notch** filter [10] with a zero at the frequency ω_0. When the pole radius R is close to the unit circle we see that ω_0 gets close to θ. That is, the pole and zero of the filter $G(z)$ are very close to each other. Thus, at frequencies sufficiently away from ω_0, the response is close to unity. This is demonstrated in Fig. 4, which shows the magnitude response of $G(z)$ for two values of R. From Eq. (2) we see that

$$\begin{bmatrix} G(e^{j\omega}) \\ H(e^{j\omega}) \end{bmatrix} = \frac{\mathbf{U}}{\sqrt{2}}\begin{bmatrix} 1 \\ A(e^{j\omega}) \end{bmatrix}$$

where $\mathbf{U}$ is unitary, that is, $\mathbf{U}^t\mathbf{U} = \mathbf{I}$. This shows that

$$|G(e^{j\omega})|^2 + |H(e^{j\omega})|^2 = \frac{1 + |A(e^{j\omega})|^2}{2} = 1$$

where we have used the property $|A(e^{j\omega})| = 1$. It therefore follows that $G(z)$ and $H(z)$ are power complementary. This shows, in particular, that the filter $H(z)$ is a good antinotch filter as demonstrated in Fig. 5, for the same pole radii chosen in Fig. 4.

By choosing $\omega_0 = 2\pi/3$ the filter $H(z)$ can be used to extract the period-3 regions of the DNA effectively. The allpass filter $A(z)$ can be implemented with either the direct form structure or the cascaded lattice structure [8], [13]. The lattice structure with one-multiplier sections [13] is especially attractive [10], and Fig. 6 shows the implementation of $H(z)$ using this lattice. The multipliers in this structure are the lattice coefficients

$$k_1 = R^2, \quad k_2 = -\cos\omega_0.$$

Since the antinotch frequency is $\omega_0 = 2\pi/3$ we have

$$k_2 = -\cos\omega_0 = 1/2$$

which can be implemented with a binary shift. So the only significant multiplier is R^2, and controls the antinotch quality without affecting the frequency ω_0 (Fig. 5). Thus we can adjust R^2 depending on the base-domain resolution desired.

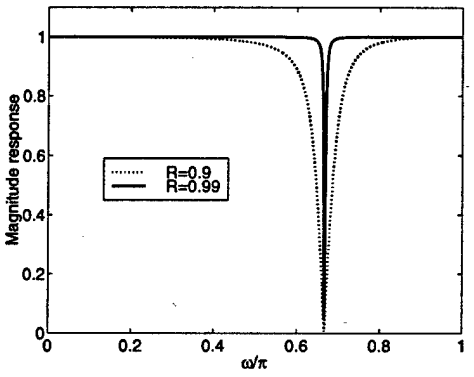

Figure 4. Notch filter responses for two values of R.

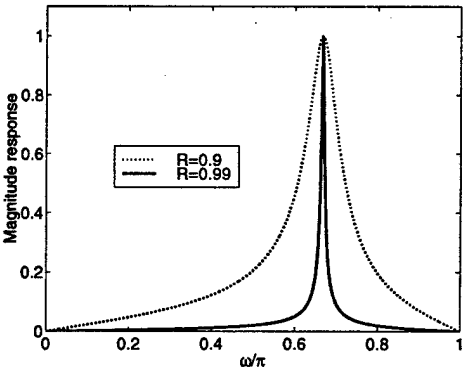

Figure 5. Antinotch filter responses for two values of R.

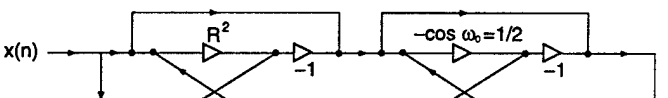

Figure 6. Lattice structure for implementing the antinotch filter $H(z) = V(z)/X(z)$.

V. MULTISTAGE FILTERS

Even though the IIR antinotch method has been found to work well [14], there is room for improvement. We will show that with a slight increase in the number of multipliers we can design filters with much better stopband attenuation. Such filters are essential in order to suppress the background $1/f$ noise which is always there in the DNAs of many organisms, due to long-range correlation between base pairs.

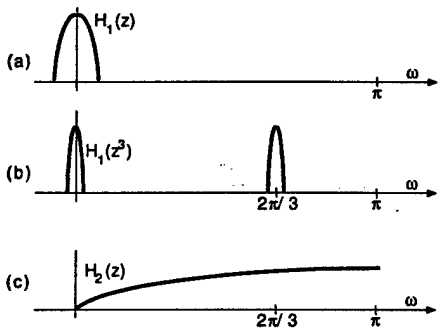

Figure 7. The multistage design of a narrowband bandpass filter. (a) Magnitude response of lowpass prototype $H_1(z)$, (b) multiband response of $H_1(z^3)$, (c) the response of $H_2(z)$ which eliminates the unwanted passband at $\omega = 0$.

The method to be presented is based on the idea of multistage filtering [3,13]. To explain this consider a narrowband lowpass filter $H_1(z)$ as shown in Fig. 7(a). If we replace each delay element z^{-1} in the filter with z^{-3}, we get the filter $H_1(z^3)$ whose response is as shown in Fig. 7(b). Thus, there is a passband centered at $2\pi/3$ and a passband at $\omega = 0$. If we now cascade this with a filter $H_2(z)$ which attenuates the zero-frequency passband severely, the resulting filter

$$H(z) = H_1(z^3)H_2(z)$$

is a narrowband filter with passband centered at $2\pi/3$. We will demonstrate that $H_1(z)$ and $H_2(z)$ can be designed with very low complexity, and that the filter predicts the exons with good accuracy. The multistage idea is similar in principle to the IFIR method introduced by Neuvo, et al. [7,13].

Figure 8 shows an example. Here $H_1(z)$ is a third order elliptic filter and $H_2(z)$ is chosen to have two zeros at $\omega = 0$, that is,

$$H_2(z) = (1 - z^{-1})^2.$$

The various filter responses involved in the multistage design are shown in the figure. The bottom plot shows the multistage filter $H(z)$ which has a narrow passband at $\omega = 2\pi/3$, and excellent attenuation at most frequencies. Implemented in direct form [8], $H_1(z)$ requires 5 multipliers, and $H_2(z)$ is multiplierless.

It should be noticed here that $H_1(z)$ can be implemented using the allpass decomposition method [13], which allows the third order elliptic filter to be written in the form

$$H_1(z) = \frac{A_0(z) + A_1(z)}{2}$$

where $A_0(z)$ is a first order allpass filter and $A_1(z)$ a second order allpass filter, both with real coefficients. We can implement $A_0(z)$ and $A_1(z)$ with one and two multipliers respectively [13], so that $H_1(z)$ requires only three multipliers. Summarizing, the multistage method has only slightly higher complexity than the allpass-based antinotch filter, but its characteristics are significantly better.

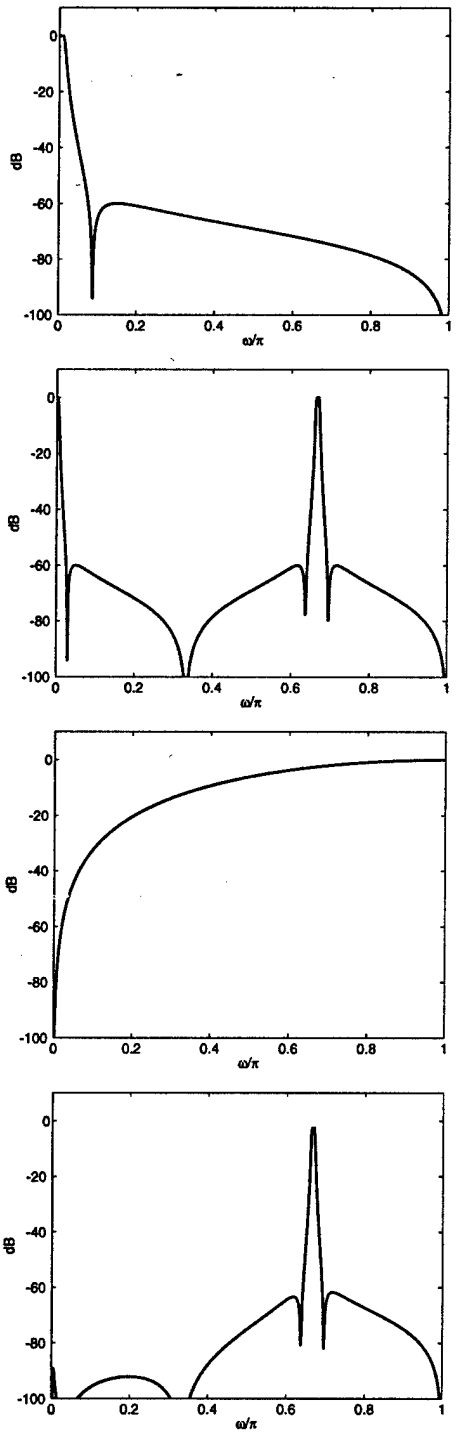

Figure 8. Magnitude responses of filters in the multistage design method. From top to bottom: the IIR lowpass filter $H_1(z)$, the expanded version $H_1(z^3)$, the FIR filter $H_2(z)$, and the multistage filter $H_1(z^3)H_2(z)$.

VI. EXAMPLES AND CONCLUSIONS

We show in Fig. 9 the exon prediction results for gene F56F11.4 in the C-elegans chromosome III. This gene has five exons. The first plot uses the DFT based spectrum described in Sec. III. The five peaks corresponding to the exons can be seen clearly. The middle plot uses the allpass-based antinotch filter with pole radius $R = 0.992$. This scheme can be implemented with only one multiplier per output sample (i.e., per base pair). Both of these methods locate the five exons quite well, but we also notice the background "noise" due to the $1/f$ characteristics in DNA sequences. The third plot uses the multistage filter $H(z)$ shown in the bottom of Fig. 8. Notice that the background noise has been removed almost completely and the five exons can be seen clearly.

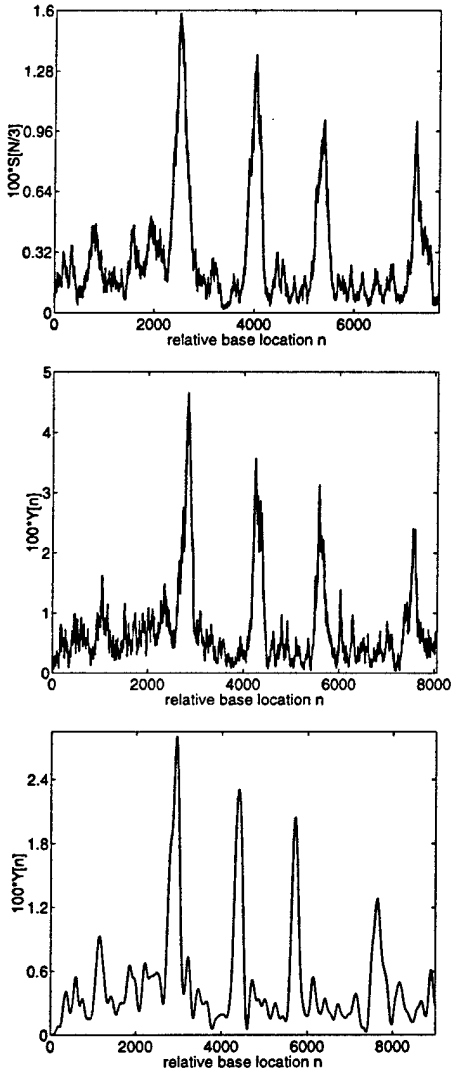

Figure 9. Top plot: the DFT based spectrum $S[N/3]$ for gene F56F11.4 in the C-elegans chromosome III. Middle plot: the antinotch filter output (Sec. IV) for the same gene. Bottom plot: the multistage narrowband bandpass filter output (Sec. V) for the same gene.

As explained in detail in [4], gene identification is a very complex problem, and the identification of period-3 regions is only a step towards gene and exon identification. In fact, Tiwari, et al. [11] have observed that some genes do not exhibit period-3 behavior at all in S. cerevisiae (e.g., genes of the mating type locus). The period-3 property has often been attributed to the dominance of the base G at certain codon positions in the coding regions. We have, in fact, observed experimentally that the use of the base G alone, instead of all four bases, often leads to excellent prediction of period-3 regions.

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] R. E. Crochiere, and L. R. Rabiner, *Multirate digital signal processing,* Prentice Hall, Inc., 1983.

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

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

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

[7] Y. Neuvo, and C.-Y. Dong, and S. K. Mitra, "Interpolated finite impulse response filters," IEEE Trans. on ASSP, pp. 563–570, June. 1984.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] P. P. Vaidyanathan, and B-J. Yoon, "Gene and exon prediction using allpass-based filters," Workshop on Genomic Sig. Proc. and Stat., Raleigh, NC, Oct. 2002.

[15] 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.