The role of signal-processing concepts in genomics and proteomics

P. P. Vaidyanathan and Byung-Jun Yoon

Contact Author:[1] P. P. Vaidyanathan,[2] Dept. Electrical Engr., 136-93, California Institute of Technology, Pasadena, CA 91125. Ph:(626) 395 4681 Email: ppvnath@systems.caltech.edu

Abstract. With the enormous amount of genomic and proteomic data that is available to us in the public domain, it is becoming increasingly important to be able to process this information in ways that are useful to humankind. Signal processing methods have played an important role in this context, some of which are reviewed in this paper. First we review the role of digital filtering techniques in gene identification. We then discuss the topic of long-range correlation between base pairs in DNA sequences. This correlation corresponds to a $1/f$ type of power spectrum. We also describe some of the recent applications of Fourier methods in the study of proteins. Finally we mention the role of Karhunen-Loeve like transforms in the interpretation of DNA microarray data for gene expression.

[1]Work supported in part by the ONR grant N00014-99-1-1002.
[2]Invited paper, Journal of the Franklin Institute, special issue on Genomics, 2004.

1. INTRODUCTION

With the enormous amount of genomic and proteomic data that is available to us in the public domain, it is becoming increasingly important to be able to process this information in ways that are useful to humankind. In this context, traditional as well as modern signal processing methods have played an important role in these fields. The reader will find a number of papers in this special volume addressing such issues. Our goal in the present paper is to concentrate on some of the areas where well-established signal processing techniques have had a role. The first one is the role of digital filtering techniques in gene identification. The idea arises from the fact that protein coding regions (exons within genes) typically exhibit a period-3 behavior that is not found in other parts of the DNA molecule. After a review of recent results on this topic in Sec. 3, we move on to another fascinating property observed in DNA sequences, namely the presence of long-range correlation between base pairs. This correlation corresponds to a $1/f$ type of power spectrum. The history behind this is reviewed in Sec. 4. In Sec. 5 we describe some of the recent applications of traditional Fourier transformation in the study of proteins (which are sequences of twenty possible amino acids). Finally in Sec. 6 we briefly review the role of Karhunen-Loeve like transforms in the interpretation of DNA microarray data for gene expression. In each of the above sections several original references are mentioned and the interested reader should pursue these in detail. We begin with a brief review of DNA-related fundamentals in Sec. 2.

2. SOME FUNDAMENTALS

An overview of some of the important aspects of the DNA molecule from the signal processing view point can be found in the introductory magazine-article by Anastassiou [4]. 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, and 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 corresponds 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). 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).

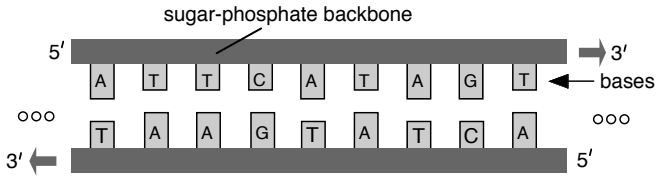

Figure 1. The DNA double helix (linearized schematic).

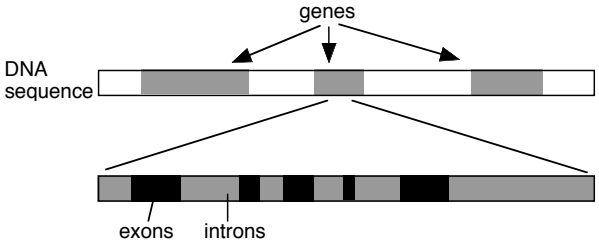

Figure 2. A DNA sequence has genes and intergenic regions. The genes of eucaryotes have exons (protein coding regions) and introns.

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 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 because they are removed in the process of forming the RNA molecules (called messenger RNA or mRNA). Thus, unlike the parent gene, the mRNA has no introns; it is a concatenation of the exons in the gene. The mRNA carries the genetic code to the protein machinery in the cell called the ribosome (located outside the nucleus). The ribosome produces the protein coded by the gene.

3. FILTERS FOR GENE IDENTIFICATION

It is well-known that base sequences in the protein-coding regions of DNA molecules have a **period-3 component** because of the codon structure involved in the translation of base sequences into amino acids This observation can be traced back to the 1980 work of Trifonov and Sussman [29]. 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 [28]. Furthermore for procaryotes (cells without a nucleus), and some viral and mitochondrial base sequences, such periodicity has even been observed in noncoding regions [17]. For this and many other reasons, gene prediction is a very complicated problem (see the review article by Fickett [10]). 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 (e.g., see Fig. 13 later). In fact this technique has also been used to isolate exons within the genes of eucaryotic cells [4, 28]. 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 (see Sec. 4).

In this section we provide an efficient mechanism for the identification of DNA regions exhibiting period-3 behavior. This is based on digital filtering methods presented earlier in [31] and [32]. 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. 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 demonstrated on gene sequences taken from C.elegans in the public genomic database.

Codon bias. Some authors have claimed that the period-3 property is due to nonuniform codon usage, also known as codon bias: even though there are several codons which could code a given amino acid, they are not used with uniform probability in organisms. This creates a codon bias. There is an excess of guanine (G) in position 1, leading to strong period 3 oscillation [13]. 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 amino acids 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

intact! Tiwari, et al. also observe that some genes do not exhibit period-3 at all in S. Cerevisiae.

3.1. DNA spectrum and 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

$$S[k] \stackrel{\Delta}{=} |X_A[k]|^2 + |X_T[k]|^2 + |X_C[k]|^2 + |X_G[k]|^2 \tag{1}$$

then we should see a peak at the sample value $k = N/3$ as demonstrated in many papers (e.g., [28]). 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 (Sec. 4). However a long window implies longer computation time, and also compromises the base-domain resolution in predicting the exon location.

3.2. Relation to filtering

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 [7, 30]. 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. 3). 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.

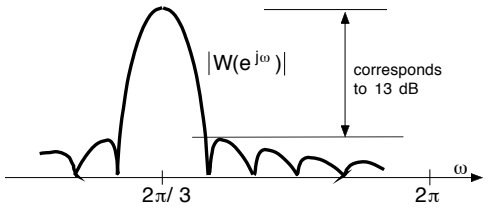

Figure 3. The filtering effect of DFT computation.

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. 4. 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.

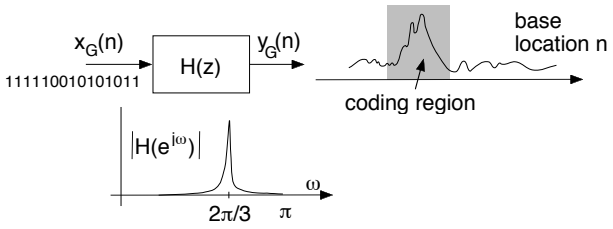

Figure 4. A digital filter $H(z)$ with indicator sequence $x_G(n)$ as its input.

3.3. IIR Antinotch Filters

The use of IIR antinotch filters for gene prediction was proposed in [31]. 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

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

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 [24] 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 (Fig. 5). Thus, at frequencies sufficiently away from ω_0, the response is close to unity. This is demonstrated in Fig. 6, which shows the magnitude response of $G(z)$ for two values of R.

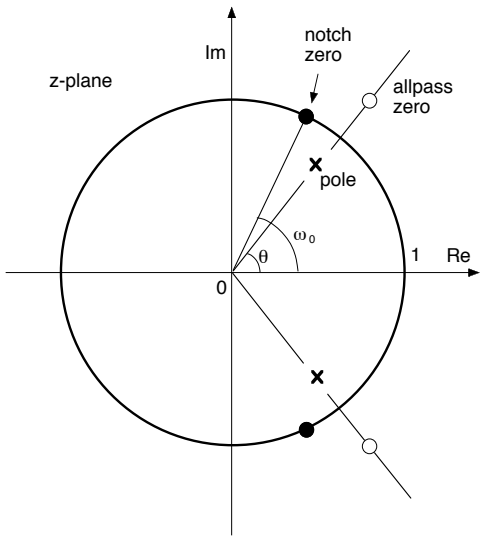

Figure 5. Poles and zeros of the notch filter $G(z)$ and the allpass filter $A(z)$.

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 allpass 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. 7, for the same pole radii chosen in Fig. 6.

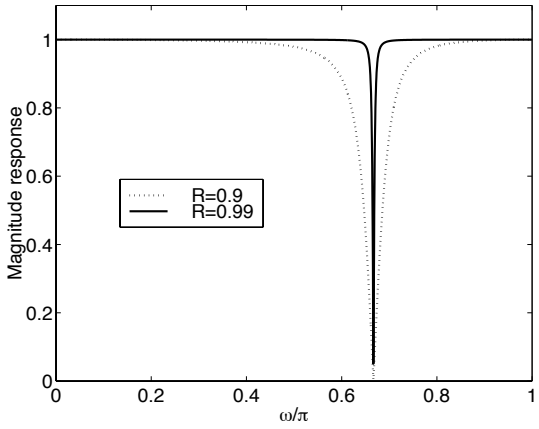

Figure 6. Notch filter responses for two values of R.

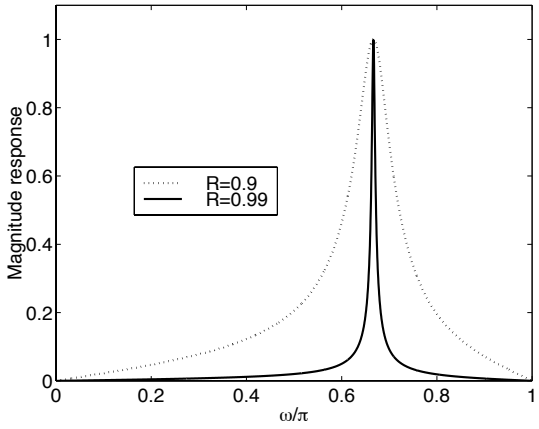

Figure 7. Antinotch filter responses for two values of R.

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 [20], [30]. The lattice structure with one-multiplier sections is especially attractive [24], [30], and

Fig. 8 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. 7). Thus we can adjust R^2 depending on the base-domain resolution desired.

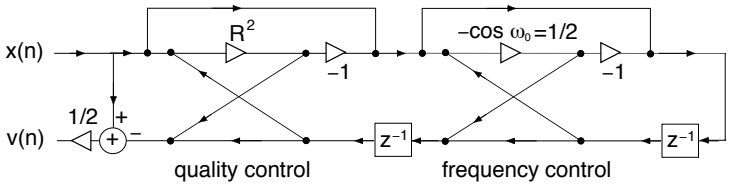

Figure 8. Lattice structure for implementing the antinotch filter $H(z) = V(z)/X(z)$.

3.4. Multistage Filters

Even though the IIR antinotch method has been found to work well, 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.

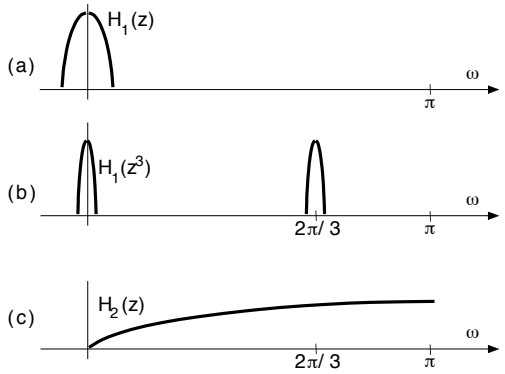

Figure 9. The multistage design of a narrow band bandpass filter. (a) Magnitude response of lowpass prototype $H_1(z)$, (b) multiband response of $H_1(z^3)$, (c) the response of $H_2(z)$ which eliminates the unwanted passband at $\omega = 0$.

The method to be presented is based on the idea of multistage filtering [7,30]. To explain this consider a narrowband lowpass filter $H_1(z)$ as shown in Fig. 9(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. 9(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 so-called IFIR method introduced by Neuvo, et al. [19,30].

Figure 10 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 [20], $H_1(z)$ 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 [30], 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 [30], 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.

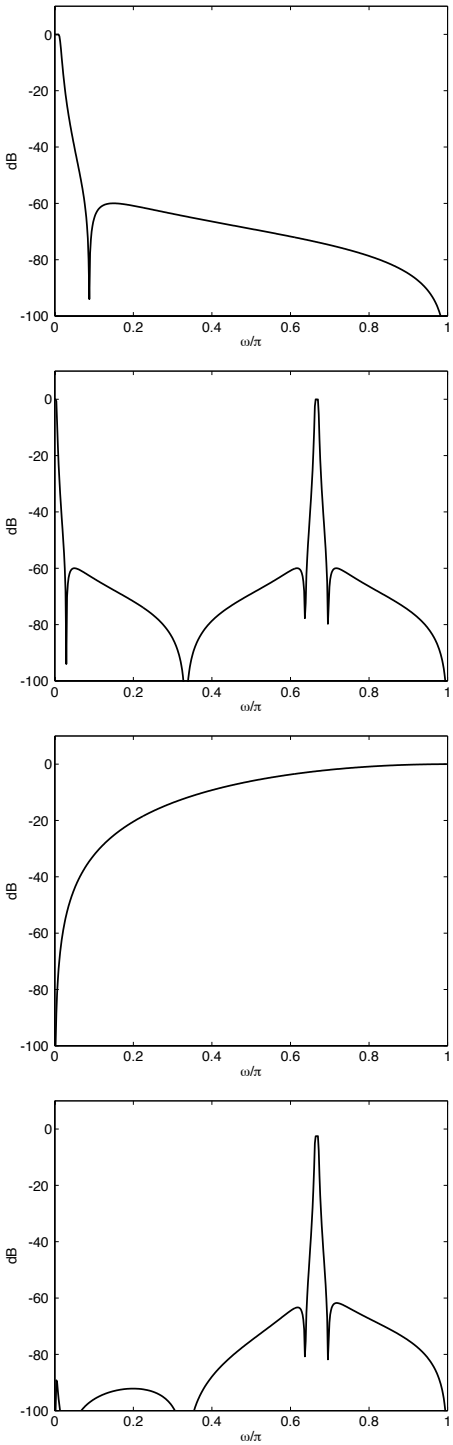

Figure 10. Magnitude responses of filters in the multistage design method. From top to bottom: the IIR lowpass filter $H_1(z)$, the expanded version $H_1(z^3)$, the FIR filter $H_2(z)$, and the multistage filter $H_1(z^3)H_2(z)$.

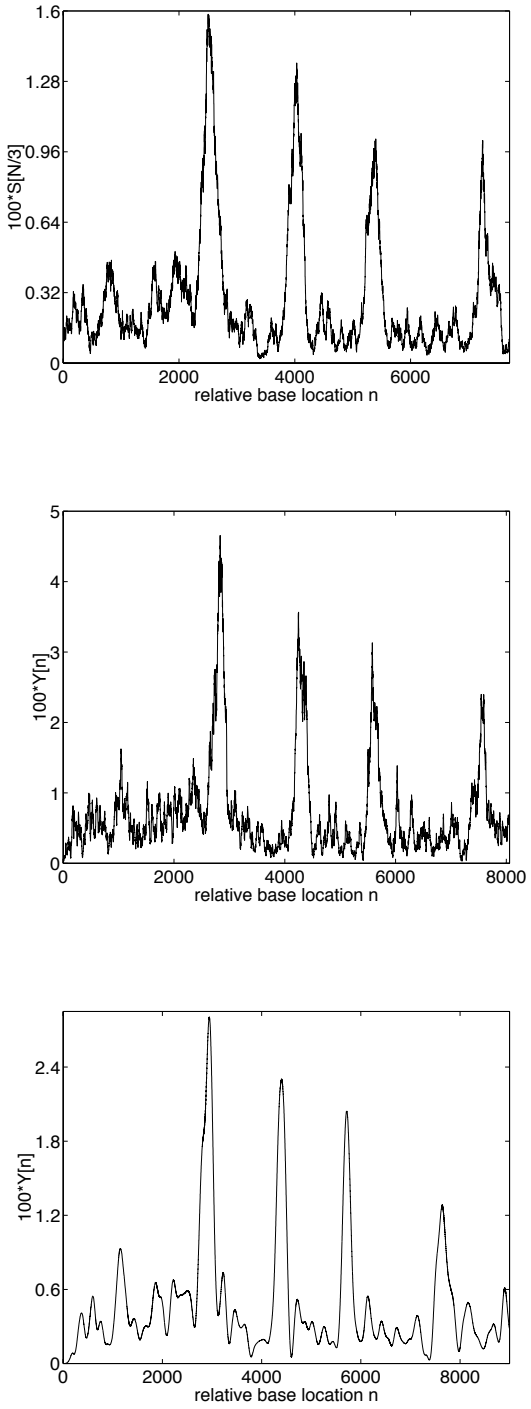

Figure 11. Top plot: the DFT based spectrum $S[N/3]$ for gene F56F11.4 in the C-elegans chromosome III. Middle plot: the antinotch filter output (Sec. 3.3) for the same gene. Bottom plot: the multistage narrowband bandpass filter output (Sec. 3.4) for the same gene.

We show in Fig. 11 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. 3.1. 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. 10. Notice that the background noise has been removed almost completely and the five exons can be seen clearly.

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. As explained in detail in [10], 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. [28] have observed that some genes do not exhibit period-3 behavior at all in S. cerevisiae (e.g., genes of the mating type locus).

4. LONG RANGE CORRELATIONS IN DNA

A curious observation about DNA squences is the fact that base pairs that are far away are still "correlated" in a statistical sense. Given the fact that DNA sequences are very long (millions of bases even in the simplest microbial organisms) this observation is quite interesting. One of the earliest papers to point out the long range correlations in DNA sequences was the 1992 paper by Peng, et. al in *Nature* [22]. These authors observed long range correlations in genes with introns, but not in intronless genes and in complementary DNA. The study was made based on a concept called the *DNA walk*. Latter studies by other authors examined correlations over much longer regions which contained many genes. More careful studies have also indicated long range correlations both in coding and noncoding regions (see, for example, [37] and references thererin).

Another early work on the topic was the 1992 paper by Richard Voss [34] who was perhaps also the first person to define indicator sequences for bases, and calculate the deterministic autocorrelation. For example, letting $x_A(n)$ be the indicator for base A, the autocorrelation is

$$r_A(k) = \sum_n x_A(n) x_A(n-k) \tag{3}$$

where the sum extends over all n for which the product is nonzero. The Fourier transform of this yields the

power-spectrum for base A:

$$S_A(e^{j\omega}) = S_A(e^{j2\pi f}) = \sum_k r_A(k) e^{-j2\pi kf} \tag{4}$$

Notice that $S_A(e^{j\omega}) = |X_A(e^{j\omega})|^2$. Voss analyzed the human Cytomegalovirus strain AD169. The genome length was $N = 229,354$. The lowest meaningful frequency[3] can be regarded as $1/N$ which is slightly smaller than $0.5 * 10^{-5}$. Voss demonstrated that the power spectrum has power-law or $1/f^\beta$ behavior for each of the four indicator sequences, even though β varied from base to base (0.76 for A, 0.77 for T, 0.83 for C and 0.92 for G). The important property that the power spectra have peaks at the frequency $f = 1/3$ was also observed by Voss and the possibility that this might be due to the codon structure was also correctly predicted! At higher frequencies the power spectrum tended to flatten out, representing white-noise like behavior. Voss also studied many other organisms categorized in the genebank (bacteria, plants, mammals, even viruses) and calculated the exponent β in the power sepectrum (averaged over many genomes in each category). For example $\beta \approx 1$ for invertebrates (exactly $1/f$ behavior) and $\beta \approx 0.7$ for organelles.[4] Thus Voss was able to make several poineering observations, even though these were based on relatively short base sequences (over 200,000) available at that time.

We know that an impulse in the Fourier transform at zero frequency represents a constant component in the time domain. Similarly the $1/f^\beta$ behavior which implies an unbounded component at $f = 0$ represents a slowly decaying term in the autocorrelation sequence. This gives rise to the term *long-range correlation*. Qualitatively speaking, bases that are far away in the DNA molecule (e.g., separated by a quarter of a million positions) still have correlation among them. Later studies have indicated that such long range correlation is valid even further, extending to several millions of bases!

Does the $1/f$ behavior of the power spectrum of a DNA molecule extend all the way up to near-zero frequencies? To answer this question, we have to perform the measurements on longer and longer DNA sequences because the lowest meaningful frequency is $f = 1/N$. In 1999 de Sousa Vieira [33] made such calculations for thirteen microbial genes with more than a million bases each (e.g, E. Coli which has over 4.6 million bases). In each case the entire DNA sequence of length N was divided into nonoverlapping subsequences of length L, the power spectrum of each subsequence calculated, and the result was averaged over all subsequences to reduce the statistical variance of the estimate. The paper by de Sousa Vieira also reported results for the case of overlapping sliding windows. For many of the organisms studied the conclusion was that the power spectrum for each type of base flattens out at very low frequencies (like $f < 10^{-6}$) instead of continuing the $1/f$ trend. The power spectrum $S(f)$ therefore had three regions as

[3]Recall that the sample spacing for the indicator sequence $x_A(n)$ is normalized to be unity, so the highest frequency π corresponds to 0.5.

[4]Note that $\beta = 0$ corresponds to white noise and $\beta = 2$ to Brownian motion.

demonstrated qualitatively in Fig. 12. The sloping straighline indicates the $1/f$-part on the log-log scale.

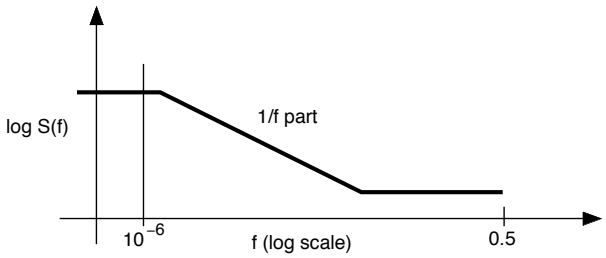

Figure 12. The three regions in the DNA spectrum. There is $1/f$ behavior except at very low and high frequencies where the spectrum flattens out.

There are always variations and exceptions. For some of the organisms (e.g., *Bacillus subtilis*) there was no flattening — the $1/f$ behavior continuing "even unto the lowest frequency $1/N$." And finally there were cases e.g., the *Haemophilus influenzae Rd*, where the power spectrum *flattened out for some bases but did not for some others!* This shows that averaging the power spectrum over all the four bases as was done in some early work is not wise.[5] In all examples de Sousa Vieira noticed that the power spectra for A and T (which pair up in the DNA double strand) were similar to each other, and so were those of C and G.

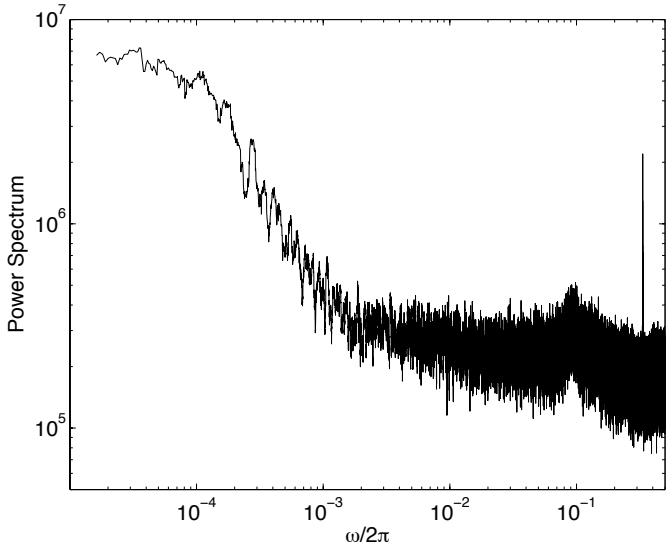

Figure 13. Demonstration of $1/f$ spectrum.

[5]The more recent paper by Sussillo, Kundaje, and Anastassiou [27] also shows clearly that different organisms indeed have significant differences in base compositions, repetetivity, and so forth.

Figure 13 shows the power spectrum $S_A(e^{j\omega})$ for base A for the first one-million bases of an entire bacterial genome of length about 1.55 million. The organism is called *Aquifex aeolicus*, and its genome can be found in the gene bank. There were 0.5 million samples of $S_A(e^{j\omega})$ in $0 \leq \omega \leq \pi$. The plot shows a slightly smoothed version with a sliding rectangular window of length 33. Notice that this is a log-log plot so the variations near zero-frequency can be seen clearly. The plot approximately resembles Fig. 12, displaying three separate regions indeed. Notice also the thin line representing a sharp peak near the right edge of the plot. This corresponds to the peak at $2\pi/3$ due to period-3 property in the coding regions (Sec. 3). Many more examples can be found in [33].

In order to compare the power spectrum of DNA sequences with random sequences, consider the following experiment. Suppose we generate a random number sequence $x(n)$ taking on four possible values $(0, 1, 2$ and $3)$ with equal probability. We can regard this as a "pseudorandom DNA sequence". Suppose we define the indicator sequences for the four values, say $x_0(n), x_1(n)$, and so forth as usual. We can compute the power spectra of these indicator sequences. Figure 14 shows an example of such a power spectrum which shows that the $1/f$ property is completely absent. The gradual thickening of the plot as f increases is an artifact of the $\log f$ axis. The same effect is also seen in Fig. 13 and should be ignored.

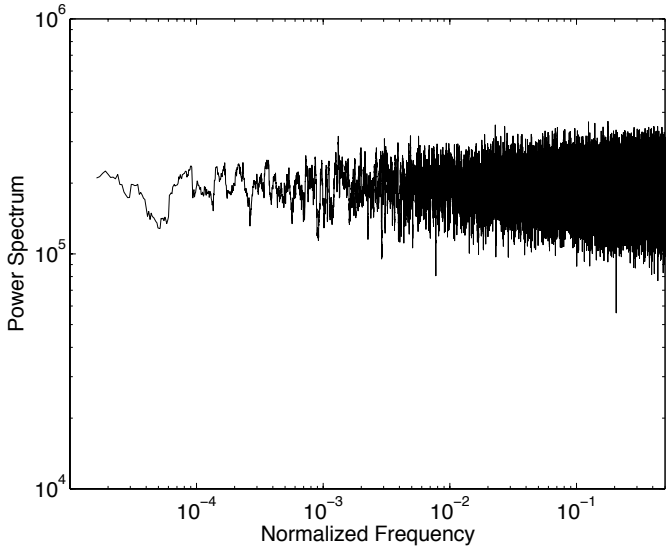

Figure 14. Power spectrum of a indicator sequence generated from a pseudorandom DNA sequence (see text). Notice the absence of $1/f$ behavior.

In an early paper [16] Li had observed that the $1/f$ behavior in natural phenomena can be traced to a certain broad general principle which he calls the expansion-modification model. When life started on earth the DNA molecules had very modest lengths (few thousand bases, perhaps even less) and as evolution progressed,

the molecules went through a lengthening processes which involved duplication and mutation. Imagine we have a character string $x_1(n)$ of length L. Suppose we duplicate it and then make some random changes of certain characters (from the same alphabet), and concatenate the result $y(n)$ to the original $x(n)$ to form a sequence that is twice as long. Suppose we repeat this process over and over again. This quickly results in a very long sequence comparable to today's DNA molecules. Furthermore it can be shown mathematically (though this is quite nontrivial [16]) that repeated application of such a duplication-mutation process results in long-range correlations such as those observed in DNA sequences. The duplication-mutation model might therefore hold the key for the $1/f$ behavior of the DNA spectrum.

Perhaps the most comprehensive work in this area as of 1997 is the excellent review paper by Li [17]. Li makes the point that long-range correlations of base sequences measure correlations at the level of too small an entity (namely individual bases). More interesting will be the study of correlations between larger units which have biological significance, rather than single bases. Li also studies all the sixteen cross-correlations between the four types of bases, e.g.,

$$r_{AG}(k) = \sum_n x_A(n) x_G(n-k) \tag{5}$$

and so forth. A curious property here is the observation that $r_{AG}(k) \approx r_{CT}(k)$ and similarly $r_{AA}(k) \approx r_{TT}(k)$ and $r_{CC}(k) \approx r_{GG}(k)$. The correlations among different types of base pairs is very interesting indeed.

The $1/f$ property is also related closely to fractal behavior as seen from the paper by Wornell [36]. Wornell has also established that a natural way to model $1/f$ processes would be to use a wavelet transform domain, or equivalently, a tree structured filter bank. Consistent with this idea, Hausdorff and Peng proposed in 1996 the so-called **multiscale randomness model** as a possible source of $1/f$ behavior in DNA sequences [12].

In addition to the overall $1/f$ behavior of DNA sequences, and the period-3 property in protein coding regions, it has been observed by many authors that DNA molecules also have components of period 10 to 11 (see [13] and references therein). In this 1998 paper Herzel, et. al have argued that this periodicity can be attributed to an alternation property in protein molecules. In these molecules the hydrophilic and hydophobic regions (water loving and water hating regions) alternate at a certain rate in the three-dimensional folded form. The authors of [13] performed experiments with pseudogenes which were derived from proteins by inverse aminoacid-codon mapping. Such pseudogenes have no introns. Furthermore they are free from codon bias (Sec. 3) if we make codon assignments with uniform probability, whenever there are multiple choices. It was found that such pseudo genes still preserve the 10-11 perodicity. This leads one to conclude that this periodicity is directly related to the hydrophilic/hydophobic alterations in proteins. This is also consistent with the observations that intron regions in DNA molecules do not exhibit the 10-11 periodicity because

they do not participate in protein coding.

There has been a very thorough study of the spectrograms of genomic sequences in the recent paper by Sussillo, Kundaje, and Anastassiou [27]. This paper shows that conclusions about the reasons for periodicites should be made with greater care. For example 10-11 periodicities have even been observed in non coding regions, and sometimes do not have anything to do with the hydrophillic/hydrophobic regions in proteins.

5. FOURIER TRANSFORMS AND PROTEIN MOLECULES

For the purposes of our discussion, a protein molecule is a long sequence of amino acids connected together by a covalent peptide bond [1]. As mentioned in Sec. 2 there are upto 20 different amino acids in the proteins of living organisms. There are innumerable combinations of such acids and the resulting number of proteins in living organisms is therefore enormous. Proteins drive most of the biological processes in living organisms. Enzymes, for example, are proteins with a special role, namely the speeding up some of the biochemical reactions in living organisms. Of fundamental importance in protein functioning is the ability of a protein to interact selectively with a small number of other molecules. This ability is derived from the fact that a protein molecule folds beautifully into a three dimensional shape determined entirely by the amino acid sequence that makes it up. The 3D shape allows certain other molecules to attach to the protein at specific sites, sometimes referred to as hot spots. A protein molecule typically has many functions (many hot spots). Given a collection of proteins, suppose they all have one function in common. Is there a mathematical way to identify this commonality simply by analyzing the amino acid sequence?

This has indeed been found to be possible based on Fourier techniques. The theory behind this, based on the so-called resonant recognition model (RRM), is described in [6]. This allows one to identify the common hot spots of many protein molecules using Fourier transform methods. This theory has later been applied [23] for the study of functional and structural relationships of a special class of proteins called **oncogene proteins** which are responsible for cancerous cell growth.[6] The power of the wavelet transform in this context has also been described in [23]. In this review we shall be content with describing the basic idea. The interested reader should really read the above references in detail.

With each amino acid molecule in a protein it is possible to associate a unique nonnegative number called the EIIP (average electron-ion interaction potential). This number ranges from 0.0 to 0.1263. The amino acids leucine (Leu) and isolucine (Ile) have the smallest value of zero, and aspartic acid (Asp) has the highest value of 0.1263. The EIIP is plotted in Fig. 15 for the twenty amino acids. The names of the acids are not

[6] In a nutshell, an oncogene is a mutated (slightly altered) form of a normal gene called the proto-oncogene. The latter is a growth regulator; it is responsible for the generation of certain crucial proteins which control cell growth. The oncogene being a mutated version does not generate the right protein in right amounts, and the result is uncontrolled cell division. See [1] for more details.

indicated because we shall not require them here. Given a protein, we can associate a numerical sequence $x(n)$ with it such that $x(n)$ is equal to the EIIP value of the nth amino acid in the protein. The argument n can be regarded as equispaced distance, with spacing T determined by the amino acid spacing (≈ 3.8 Å).

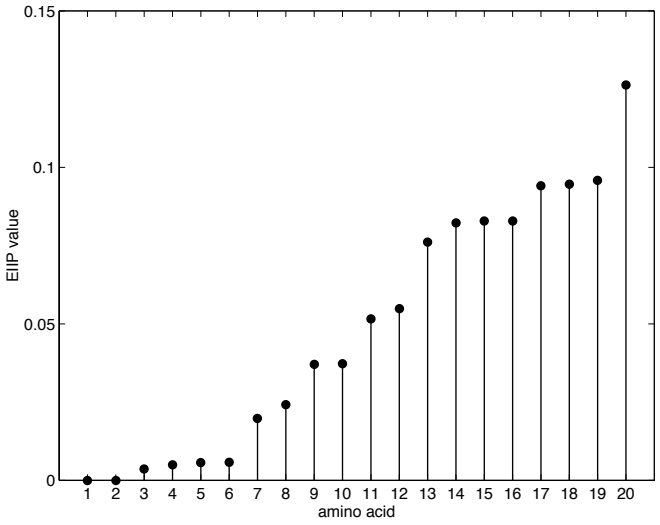

Figure 15. Plot of the electron-ion interaction potential (EIIP) for the twenty amino acids.

We can calculate the Fourier transform of $x(n)$ in the usual way

$$X(e^{j\omega}) = \sum_{n=0}^{N-1} x(n)e^{-j\omega n} \tag{6}$$

where N is the length of the amino acid sequence determining the protein. Usually a plot of $|X(e^{j\omega})|$ does not reveal much (e.g., see top plot in Fig. 17), but assume that we have a **group** of proteins. Each protein may have several biological functions but assume that there are some functions that are common to all these proteins. Define the magnitude of the product of the Fourier transforms associated with these proteins as follows:

$$P(e^{j\omega}) = |X_1(e^{j\omega})X_2(e^{j\omega})\dots X_M(e^{j\omega})| \tag{7}$$

It turns out that this product reveals an interesting feature about *biological functions that are common to this group of M proteins*. To be more precise, it has been observed through extensive experiments that if a group of proteins has only one common function then the product spectrum $P(e^{j\omega})$ has one significant peak (Fig. 16). The product $P(e^{j\omega})$ has been referred to as the **consensus spectrum** among the group of proteins used in its definition.

Interestingly enough, it has been verified experimentally [6] that peak frequencies are different for different

functions. They are therefore referred to as the **characteristic frequencies** associated with various protein function. An example is presented in [23] which considers a group of 46 oncogenes and the associated proteins (28 viral proteins, and 18 cellular proteins). The authors have observed that the consensus spectrum has precisely one peak at $\omega = 2\pi f_0$ where $f_0 \approx 0.0322$. The common functionality of these oncogene proteins is claimed to be their "ability to transform cells" [23].

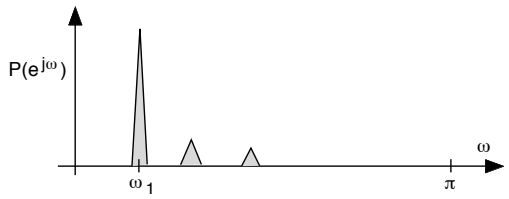

Figure 16. The consensus spectrum of a group of proteins shows the characteristic frequency of the common function of the group.

In [6] the author also demonstrates the Fourier transforms of the EIIP sequences for two proteins called FGF basic bovine and FGF acidic bovine. The amino acid chains have lengths 146 and 140 respectively. The Fourier transforms of the two EIIP sequences and the consensus spectra are shown in Fig. 17 (the plot shows magnitude-squares). Notice that the product spectrum has a single sharp peak even though the individual spectra appear to have local peaks all over the place!

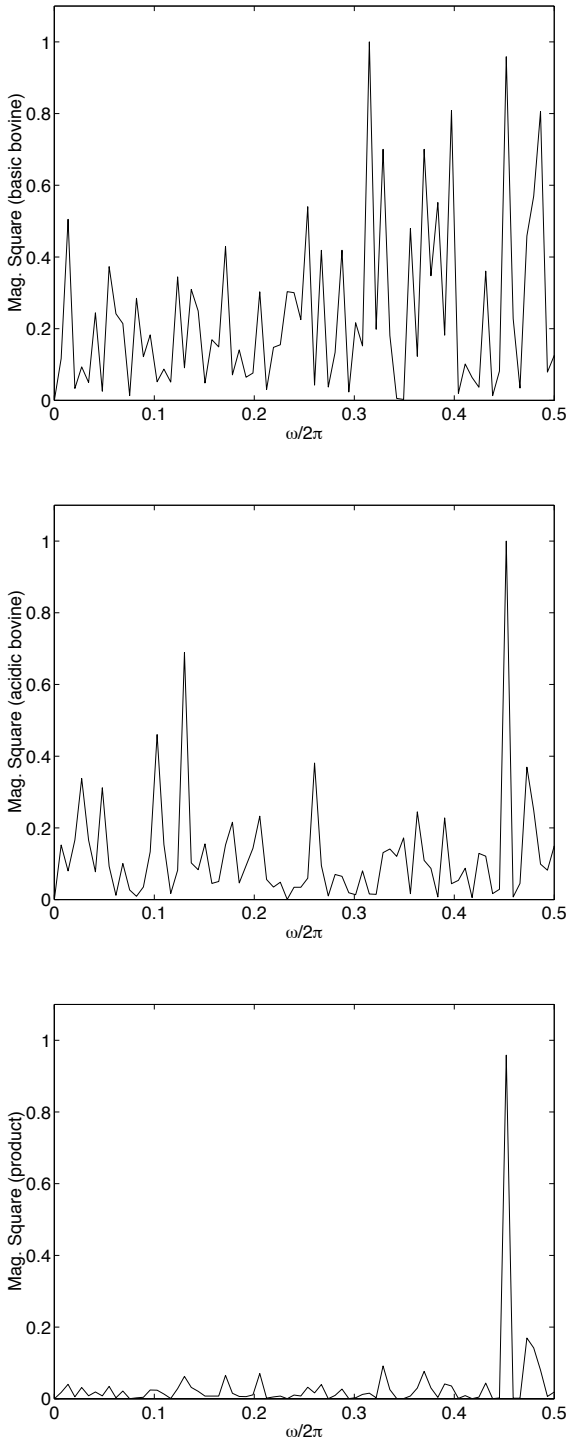

Figure 17. Magnitude squares of the Fourier transforms of the EIIP sequences for the proteins FGF basic bovine (top) and FGF acidic bovine (middle). The product, which represents the square of the consensus spectrum, is plotted in the bottom [6].

Suppose we have identified that a certain function of a protein is associated with the characteristic frequency ω_1. Is it possible to identify the amino acids that are primarily responsible for that function (i.e., identify the hot spots in the 3D protein structure which are responsible for one particular function)? This is tricky because the value of a Fourier transform at a given frequecy depends on the time domain signal at all values of time. Similarly, $P(e^{j\omega})$ depends on the EIIP values of all the amino acids. The authors of [23] propose a simple way to address this time-domain (amino-acid domain) localization, and the interested reader should read [23]. A second method would be to use the wavelet transform which gives a complete time-frequency picture, from which we can get the desired localization information. This method is also reported in [23].[7] One advantage of being able to identify a characteristic frequency with a particular funtionality is that it is then possible to synthesize artificial amino acid sequences or peptides[8] to perform certain specific functions. These could be potentially useful in drug design.

The correct functioning of a protein molecule depends upon its ability to attach selectively to certain target molecules. This is made possible by the specific three-dimensional structure of the protein. It is observed in [6] that a protein and its target molecule share a common characterisitc frequency. The name *resonant recognition model* stems from this, because some kind of resonance is suggested in the recognition of one molecule by the other. Is there a physical basis for the existence of resonant frequencies in proteins? It is suggested in [6] that there could actually be energy transfer at the visible and infrared ranges (which are the typical ranges of the characteristic frequencies, once we take into account the fact that one unit of the integer index n in our EIIP sequence $x(n)$ corresponds to 3.8 Å in space). We refrain from further elaboration on this complicated issue which is perhaps best left to the biophysicist.

6. SIGNAL PROCESSING OF MICRO ARRAY DATA

The DNA microarray records the gene expression levels of several genes on a slide.[9] The technique can be used to study simultaneously the expression levels of many genes as a function of time in a cell cycle. This offers a great deal of information to biologists. Several articles appear in a special issue of the *Nature* (e.g., see [5]) dedicated to the topic of DNA arrays. A number of interesting signal processing issues are involved in analyzing the data recorded on a DNA microarray. This includes normalization [35], data clustering, denoising, and data interpretation by linear transformations. In this context the work by Alter, et. al [2] is especially interesting. In that work, a two dimensional array or matrix

$$\mathbf{X} = [x_{nm}], \quad 0 \leq n \leq N-1,\ 0 \leq m \leq M-1 \tag{8}$$

[7] A detailed study of the use of wavelet transforms in protein structures can be found in the recent paper by Murray, Gorse, and Thornton [18].

[8] Peptides are amino acid sequences that are relatively small compared to commonly encountered proteins.

[9] The gene expression level is typically measured by measuring the level of messenger RNA (mRNA) in cells.

is generated from the expression levels of N genes at M different occasions (e.g., M points of time in the cell cycle). Each column of this matrix represents the expression levels of the N genes at the mth occasion. Each row of the matrix on the other hand represents the expression levels of *a particular gene* at different times. It is demonstrated that a singular value decomposition (SVD) of the matrix $\mathbf{X}$ reveals information that is difficult to obtain directly by observation of the data $\mathbf{X}$. Given any real $N \times M$ matrix $\mathbf{X}$ the SVD takes the form

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T \tag{9}$$

where $\mathbf{U}$ and $\mathbf{V}$ are real unitary matrices, that is,

$$\mathbf{U}^T\mathbf{U} = \mathbf{I}_N, \quad \mathbf{V}^T\mathbf{V} = \mathbf{I}_M \tag{10}$$

and $\mathbf{\Lambda}$ is a $N \times M$ matrix with zero entries everywhere except on the diagonal. Assuming $N > M$ (which is the case in the DNA microarray application), $\mathbf{\Lambda}$ takes the form

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Sigma} \\ \mathbf{0} \end{bmatrix} \tag{11}$$

where $\mathbf{\Sigma}$ is a $M \times M$ diagonal matrix whose diagonal elements are the singular values of $\mathbf{X}$, typically arranged in the order

$$\sigma_0 \geq \sigma_1 \geq \ldots \geq \sigma_{M-1} \geq 0. \tag{12}$$

The microarray data can therefore be rewritten in the form

$$\mathbf{X} = \sigma_0\mathbf{u}_0\mathbf{v}_0^T + \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \ldots + \sigma_{M-1}\mathbf{u}_{M-1}\mathbf{v}_{M-1}^T \tag{13}$$

The vectors $\mathbf{u}_k$ are the columns of $\mathbf{U}$ and are called **eigenarrays**. The vetors $\mathbf{v}_k^T$ which are rows of $\mathbf{V}$ are called **eigengenes**. To understand how Eq. (13) should be interpreted, consider the nth row of $\mathbf{X}$ which represents the expression levels of the nth gene. This row can be written as

$$\mathbf{x}_{n,row}^T = c_{n0}\mathbf{v}_0^T + c_{n1}\mathbf{v}_1^T + \ldots + c_{n,M-1}\mathbf{v}_{M-1}^T \tag{14}$$

That is, the vector representing the nth gene's expression levels has been expressed as a linear combination of M eigengenes. Notice that all the N original genes have been represented here using $M < N$ eigengenes. Similarly, the mth column of the microarray data $\mathbf{X}$ can be expressed as

$$\mathbf{x}_{m,col} = d_{0m}\mathbf{u}_0 + d_{1m}\mathbf{u}_1 + \ldots + d_{M-1,m}\mathbf{u}_{M-1}$$

That is, the gene expressions for all N genes at a particular instance can be expressed as a linear combination of $M < N$ eigenarrays. It has been found in [2] that the eigengene vectors corresponding to the dominant

singular values σ_0 and σ_1 closely approximate sines and cosines, and it is argued therein that eigengenes can be regarded as fundamental in capturing the gene expression as a function of time. This has been found to be especially useful in deriving a mathematical basis for the traveling wave pattern that is typically observed in the microarray data of gene expression with genes appropriately ordered (e.g., see Fig. 6 in [2], and Fig. 4 in [3]). A great deal of interesting detail can be obtained from the original reference [2]. The analysis of two different DNA microarrays by simultaneous diagonalization is described in a latter paper [3].

7. CONCLUDING REMARKS

The role of signal processing in genomics and more generally biological sciences has been quite impressive. In this paper we reviewed a number of these but there are more areas that we did not touch upon. One example is the interesting role played by hidden Markov models (HMM) in gene prediction. A great deal has been written about this topic, and we refer the interested reader to the original literature, e.g., [15], [25] and references therein. Finally the role of signal processing in *DNA sequencing* cannot be under estimated either. Some of the interesting papers in this area include that of Huang, et. al [14], Davies, et. al, [8], and Zhang and Allison [38].

In recent years it has been found that there exist many genes which do not code for proteins, but instead merely genetate RNA molecules for various functions. While the roles of RNA molecules such as mRNA (which is translated into protein) and tRNA have been well known, it has been found that there are several other types of RNA molecules not translated into proteins. Examples include the siRNA (small interfering RNA), miRNA (micro RNA), and so forth. These RNAs are generated by genes, but do not eventually result in proteins; instead, they have their own functions. RNA molecules which do not generate proteins are called non coding RNA (ncRNA), and the genes which generate them are called ncRNA-genes. The estimated number of protein coding human genes is thought to be somewhere around 40,000. If the ncRNA genes are also counted, the total could be as high as 60,000. There are many recent articles on this topic; a good starting point would be the review paper by Storz [26], the Scientific American article by Gibbs [11], and the paper on identification of ncRNA-genes by Eddy [9]. These ncRNA genes do not have the period-3 structure described in Sec. 3. Identification of ncRNA-genes by computational methods is one of the most challenging problems in computational biology today.

Acknowledgements. The authors are grateful to Dr. Dan Fuhrmann for the invitation to write this article. Thanks also to David Sussillo who made many important remarks about the paper, and to Prof. Dimitris Anastassiou whose paper in the IEEE Signal Processing Magazine was responsible for triggering our interest in genomic signal processing.

8. 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] O. Alter, P. O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling", Proc. of the Natl. Acad. of Sci., USA, vol. 97, no. 18, pp. 10101–10106, Aug. 2000.

[3] O. Alter, P. O. Brown, and D. Botstein, "Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms," Proc. of the Natl. Acad. of Sci., USA, vol. 100, no. 6, pp. 3351–3356, March 2003.

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

[5] P. O. Brown and D. Botstein, "Exploring the new world of the genome with DNA microarrays", Nature America (Genetics supplement), vol. 21, pp. 33–37, Jan. 1999.

[6] I. Cosic, "Macromolecular bioactivity: is it resonant interaction between macromolecules? — theory and applications", IEEE. Trans. Biomedical Engr., vol. 41, no. 12, pp. 1101–1114, Dec. 1994.

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

[8] S. W. Davies, M. Eizenman, and S. Pasupathy, "Optimal structure for automatic processing of DNA sequences," IEEE Trans. on Biomedical Engr., vol. 46, no. 9, pp. 1044–1056, Sept. 1999.

[9] S. R. Eddy, "Computational genomics of noncoding RNA genes," Cell, vol. 109, pp. 137–140, April 2002.

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

[11] W. W. Gibbs, "The unseen genome: gems among the junk," Scientific American, pp. 46–53, Nov. 2003.

[12] H. Hausdorff and C.-K. Peng, "Multiscaled randomness: a possible source of $1/f$ noise in biology," Physical review E, vol. 54, no, 2, pp. 2154–2157, August 1996.

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

[14] W. Huang, D. R. Fuhrmann, D. G. Politte, L. J. Thomas, and D. J. States, "Filter matrix estimation in automated DNA sequencing," IEEE Trans. on Biomedical Engr., vol. 45, no. 4, pp. 422–428, April 1998.

[15] A. Krogh, I. Saira Mian, and D. Haussler, "A hidden Markov model that finds genes in E. Coli DNA", Nucleic Acids Research, vol. 22 pp. 4768–4778, 1994.

[16] W. Li, "Expansion-modification systems: A model for spatial $1/f$ spectra", Physical review A, The American Physical Society, vol. 43, no. 10, pp. 5240–5260, May, 1991.

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

[18] K. B. Murray, D. Gorse, and J. M. Thornton, "Wavelet transforms for the characterization and detection of repeating motifs," J. Molecular Biology, vol. 316, pp. 341–363, 2002.

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

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

[21] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals and systems,* Prentice Hall, Inc., 1997.

[22] 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.

[23] E. Pirogova, Q. Fang, M. Akay, and I. Cosic, "Investigation of the structural and functional relationships of oncogene proteins", Proc. of the IEEE, vol. 90, no. 12, pp. 1859–1867, Dec. 2002.

[24] 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.

[25] S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White, "Microbial gene identification using interpolated Markov models," Nucleic Acids Research, vol. 26, no. 2, pp. 544–548, 1998.

[26] G. Storz, "An expanding universe of noncoding RNAs," Science, vol. 296, pp. 1260–1263, May 2002.

[27] D. Sussillo, A. Kundaje, and D. Anastassiou, "Spectrogram analysis of genomes", Eurasip, 2003 (to appear).

[28] 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.

[29] 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.

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

[31] 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.

[32] P. P. Vaidyanathan, and B-J. Yoon, "Digital filters for gene prediction applications," IEEE Asilomar

Conference on Signals, Systems, and Computers, Monterey, CA, Nov. 2002.

[33] M. de Sousa Vieira, "Statistics of DNA sequences: a low-frequency analysis," Physical Review E, The American Physical Society, vol. 60, no. 5, pp. 5932–5937, Nov. 1999.

[34] 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.

[35] Y. Wang, J. Lu, R. Lee, Z. Gu, and R. Clarke, "Iterative normalization of cDNA microarray data," IEEE Trans. on Information Tech. in Biomedicine, vol. 6, no. 1, pp. 29–37, March 2002.

[36] G. W. Wornell, "A Karhunen-Loeve-like expansion for $1/f$ processes via wavelets," IEEE Trans. on Information Theory, vol. 36, no. 4, pp. 859–861, July 1990.

[37] Z-G. Yu, V. V. Anh, and B. Wang, "Correlation property of length sequences based on global structure of the complete genome", Physical review E, The American Physical Society, vol. 63, pp. 011903-1—011903-8, 2000.

[38] X-P. Zhang, and D. Allison, "Iterative deconvolution for automatic base scaling of the DNA electrophoresis time series," Workshop on Genomic Sig. Proc. and Stat., Raleigh, NC, Oct. 2002.