Military Technical College
Kobry El-Kobbah,
Cairo, Egypt

7th International Conference
on Electrical Engineering
ICEENG 2010

Applications of Signal Processing in Genomic Research

By

Ashraf M. Aziz *

Abstract:

There is an enormous amount of genomic data available for researcher in public databases. The genomic has the form of deoxyribonucleic acid (DNA) and plays a vital role in the function of every living cell. Hence, there is an essential need to understand the organization and the functionality of the DNA regions. Determining the key regions in DNA data is a very important problem for biologists. In order to address this important issue, various methods have been proposed from diverse disciplines such as biology, chemistry, physics, computer and electrical engineering. Signal processing theory is becoming increasingly very important in the study of genes in the field of bioinformatics. As the genomic information is digital, it can be represented in the form of numerical sequences that can be analyzed so that the results obtained are beneficial to humankind. In this paper, a quick review of molecular biology, DNA structure, followed by a brief review of signal processing methods for identifying protein-coding regions are presented.

Keywords:

DNA, protein-coding regions, genomic signal processing

* Egyptian Armed Forces, amaziz64@yahoo.com

1. Introduction:

Genomic signal processing is the analysis, processing, and the use of genomic signals for gaining biological knowledge and translating this knowledge into systems-based applications. The term bioinformatics is the use of computational techniques in the studies of genomes [1]. It is a rapidly developing area of computer science devoted to collecting, organizing, and analyzing DNA and protein sequences. The field of bioinformatics is dealt primarily with biological data encoded in digital symbol sequences, such as DNA and amino acid sequences. The high variability and the high complexity of genetic signals call for sophisticated mathematical modeling, signal processing, and information extraction methods [2]. The discovery of the double helix structure of the DNA (Deoxyribonucleic acid) molecule in 1953, by Watson and Crick [3], is one of the landmarks of molecular biology. Subsequent to this sensational discovery, there has been phenomenal progress in genomics. With the enormous amount of genomic data 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. Genomic signal processing methods have played an important role in this field.

DNA carries the genetic instructions for life, coding for proteins that we depend on to survive as well as passing on characteristics from one generation to the next. That is why DNA is known as the "secret of life" or the "building blocks of life". All living things are made of DNA; plants, animals and humans, and it is the small differences in the DNA code that makes us different from one another and makes species different from each other. The entire set of DNA is the genome of the organism. Coded in the DNA are instructions necessary for a cell's proper functioning. Those instructions are stored in specific units called genes. When a particular instruction becomes active, the corresponding gene is said to turn on or be expressed. Following the expression of a particular gene, the corresponding section of the DNA strand is copied into a less stable molecule called messenger ribo nucleic acid (mRNA). The process of producing mRNA is called transcription. The mRNA is then transferred to the ribosome, where the protein molecule is produced by interpreting the instruction in mRNA. This process of producing proteins is referred to as translation. Translation takes place according to the genetic code, which maps successive triplets of RNA bases to amino acids. Thus, a protein is a chain of amino acid units. The translation from mRNA to protein is aided by molecules called the transfer RNA (tRNA) molecules. Proteins can carry out a number of tasks, such as catalyzing reactions, transporting oxygen, regulating the production of other proteins, and many others. In summary, the way proteins are encoded by genes involves the two major steps: transcription and translation.

The major goals of genomic signal processing research are: (1) Sequencing and comparison of genomes of different species, (2) Identifying genes and determining the functions of proteins they encode, (3) Predicting the structural features of the protein from the amino acid sequence, (4) Understanding gene expression and gene protein interaction to control cellular processes, (5) Tracing the evolutionary relationships among existing species and constructing phylogenetic trees and (6) Discovering associations between gene mutations and disease.

The purpose of this paper is to present a quick review of molecular biology and a brief review of the applications of signal processing theory in the field of bioinformatics. The remainder of this paper is organized as follows. A brief outline of the basic biology and structure of DNA are discussed in Section 2. Methods for identifying protein-coding regions, based on genomic signal processing, are presented in Section 3. Section 4 contains conclusion.

2. Structure and Basic Biology of DNA:

The structure and basic biology of DNA are helpful in understanding the DNA and its function. DNA molecule has a form of a double helix, as shown in Fig. 1, with each helix represented by a sequence composed of four nucleotides: adenine (A), thymine (T), guanine (G), cytosine (C). Between the two strands of the backbone which is outside, there are the four pairs of nitrogenous bases. The backbone is a very regular structure made from sugar (deoxyribose) and phosphate. The sugar has five carbon atoms that are typically numbered from 1' to 5'. The phosphate is attached to the 5' carbon atom, whereas the base is attached to the 1' carbon. The 3' carbon also has a hydroxyl group (OH) attached to it (see Fig. 2). The two helices are bonded to each other with a hydrogen bond between each nucleotide pair. The only bindings possible between the two helices, due to the physical shape of nucleotide molecules, are A-T and G-C. Hence, if we know the sequence of one of the helices, we also know the sequence of the other. For example, if a portion of one of the DNA helices is AAGATCC, the corresponding portion of the other helix of that DNA is TTCTAGC. The function of the DNA is determined mainly by the sequence of the nucleotides from which it is composed, which is why methods for DNA analysis are usually termed ’DNA sequence analysis’ methods. Due to the one-to-one correspondence between the sequences of the two helices, as described above, most of these methods analyze the sequence of only one of the helices.

The nucleotides (A,T,G,C) can be represented by a string of characters, for example .... AGGTAC CCAAGTATAAGAAGTTA..... It is seen that life is governed by quarternary codes (nucleotides). These nucleotides are made from carbon, nitrogen, hydrogen and oxygen atoms (see Fig. 3).

There are about three billion of the nucleotides in the DNA of a single human cell. If the genome sequence corresponding to the top strand of the DNA molecule shown in Fig. 1, is AGACTGAA, thus the bottom strand in Fig. 1 is TCTGACTT which is the complement of the top strand. These base-pairings occur through hydrogen bonds [4-8]. The RNA (ribo nucleic acid) molecule is closely related to the DNA. It is also made of four bases but instead of thymine, a molecule called Uralic (U) is used. The molecule U pairs with A by hydrogen bonding just like T pairs with A. RNA molecules are short (and short-lived) single-stranded molecules which are used by the cell as temporary copies of portions of DNA. The RNA that is used to make proteins is called messenger RNA (mRNA). As shown in Fig. 4, a DNA sequence can be separated into genes and intergenic spaces. Genes contain the information for generation of proteins. Each gene is responsible for the production of a different protein. Even though all the cells in an organism have identical genes, only a selected sub subset is active in any particular family of cells.

Figure 5 shows the steps involved in the production of a protein from a gene. It is worth noting that a gene has two types of subregions called the exons (protein-coding regions) and introns. The gene is first copied into a single stranded chain called the messenger RNA or mRNA molecule. This process is called transcription.

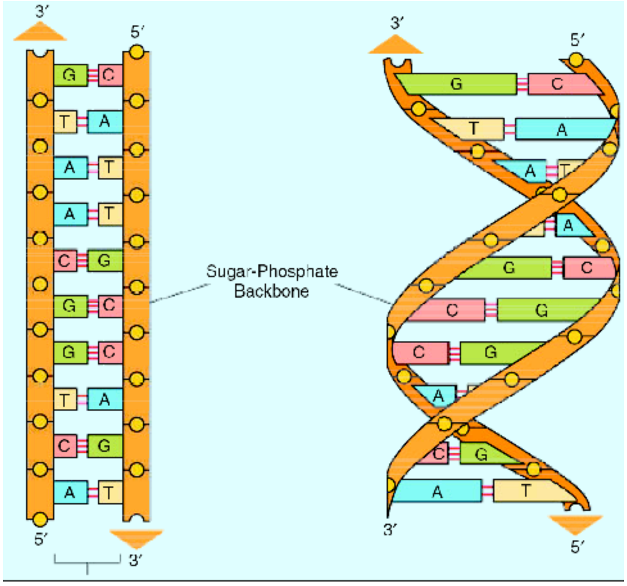


Figure (1): *DNA Structure*

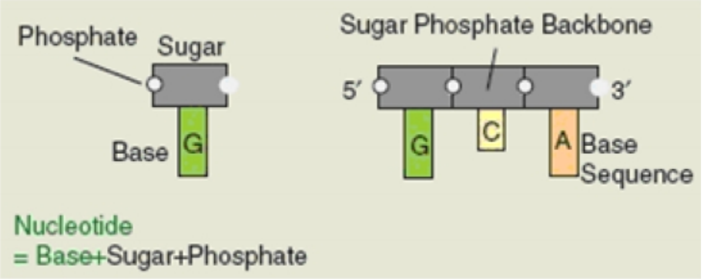


Figure (2): *Sugar Phosphate backbone*

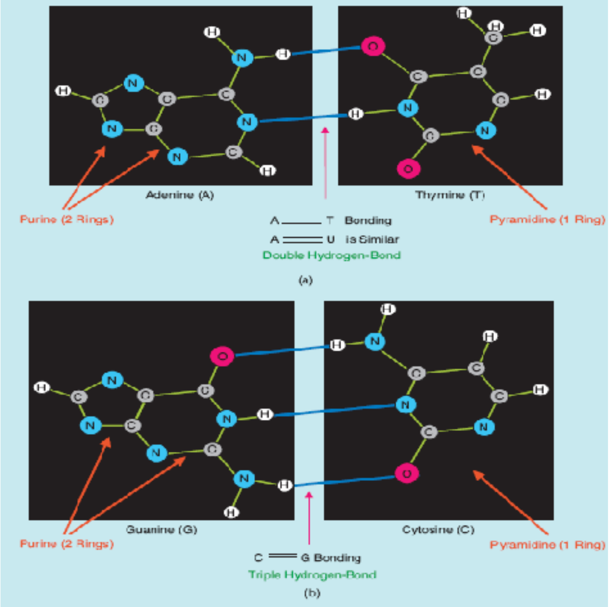


Figure (3): *Structure of nucleotides*

Figure (4): *A DNA sequence is separated into Genes and intergenic spaces*

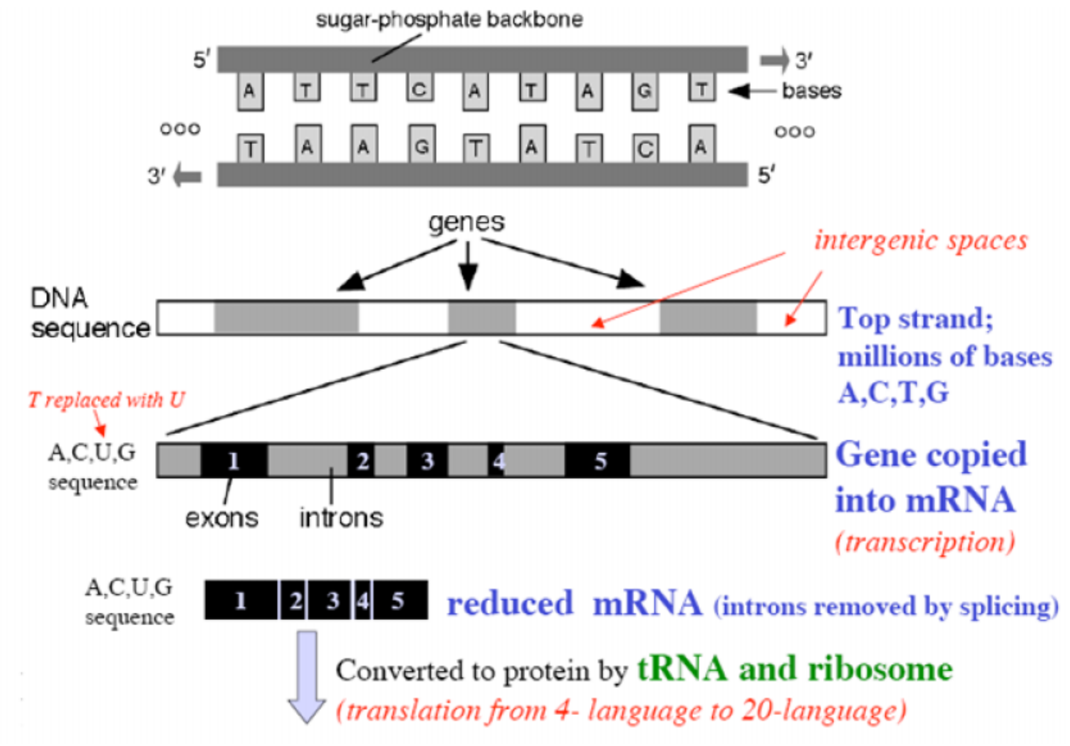
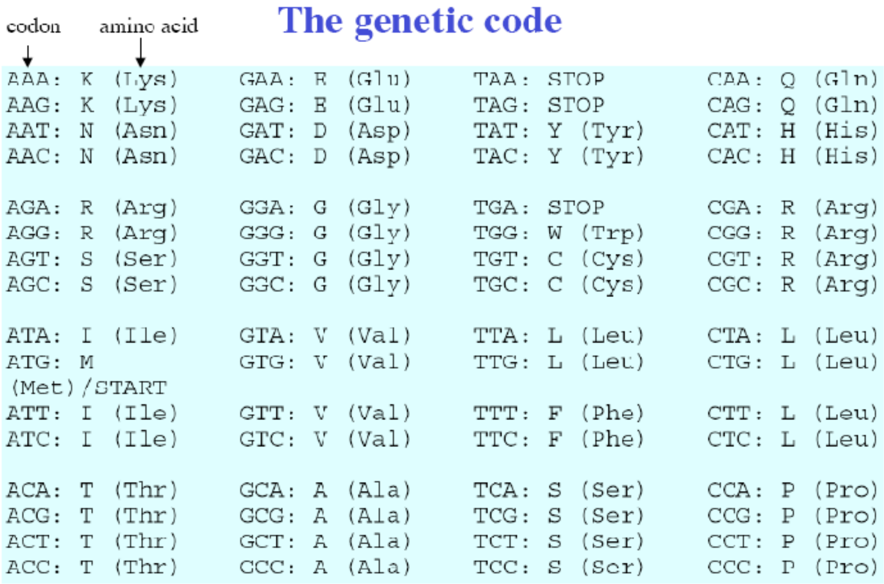


Figure (5): *The steps involved in the production of a protein from a gene*

The introns are then removed from the mRNA by a process called splicing. The spliced mRNA is then used by a large molecule (ribosome), by a process called translation, to produce the appropriate protein. When the mRNA molecule is spliced, it contains only the exons of the gene. i.e. the introns are removed. The mRNA is in reality the complement of the gene, that is, C's are replaced with G's, and A's with T's (rather U's). Thus, if the gene is AATTAGC then the mRNA is UUAAUCG. The observation that each gene creates a protein, through mRNA, is often expressed as gene in DNA→ RNA→ protein. The spliced mRNA is divided into groups of three adjacent bases.

The ribosomes move along the mRNA in 5'→3’ direction and read the mRNA sequence in nonoverlapping chunks of three nucleotides. Thus the mRNA is considered as a sequence of triplet codons. Each triplet codon instructs the cell machinery to synthesize an amino acid (codes for a particular amino acid). The codon sequence therefore uniquely identifies an amino acid sequence which defines a protein (Fig. 6).



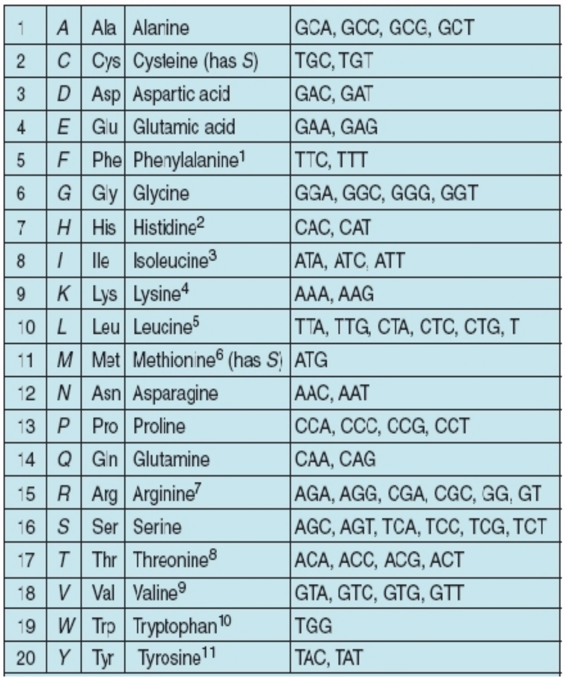
codon amino acid

The genetic code

AAA: K (Lys)	GAA: E (Glu)	TAA: STOP	CAA: Q (Gln)
AAG: K (Lys)	GAG: E (Glu)	TAG: STOP	CAG: Q (Gln)
AAT: N (Asn)	GAT: D (Asp)	TAT: Y (Tyr)	CAT: H (His)
AAC: N (Asn)	GAC: D (Asp)	TAC: Y (Tyr)	CAC: H (His)
AGA: R (Arg)	GGA: G (Gly)	TGA: STOP	CGA: R (Arg)
AGG: R (Arg)	GGG: G (Gly)	TGG: W (Trp)	CGG: R (Arg)
AGT: S (Ser)	GGT: G (Gly)	TGT: C (Cys)	CGT: R (Arg)
AGC: S (Ser)	GGC: G (Gly)	TGC: C (Cys)	CGC: R (Arg)
ATA: I (Ile)	GTA: V (Val)	TTA: L (Leu)	CTA: L (Leu)
ATG: M	GTG: V (Val)	TTG: L (Leu)	CTG: L (Leu)
(Met)/START			
ATT: I (Ile)	GTT: V (Val)	TTT: F (Phe)	CTT: L (Leu)
ATC: I (Ile)	GTC: V (Val)	TTC: F (Phe)	CTC: L (Leu)
ACA: T (Thr)	GCA: A (Ala)	TCA: S (Ser)	CCA: P (Pro)
ACG: T (Thr)	GCG: A (Ala)	TCG: S (Ser)	CCG: P (Pro)
ACT: T (Thr)	GCT: A (Ala)	TCT: S (Ser)	CCT: P (Pro)
ACC: T (Thr)	CCC: A (Ala)	TCC: S (Scr)	CCC: P (Pro)

Figure (6): *Genetic code*

As there are 4 different possible nucleotides, there are 4^3 possible codons. There are 20 types of standard amino acids that are regularly found in nature as well nonstandard types that rarely appear. Since there are only 20 possible amino acids and 64 possible codons, most amino acids can be specified by more than one triplet. i.e., the mapping from codons to amino acids is many-to one (Fig.7). For example, each of the triplets GCA, GCC, GCG, and GCT represents the amino acid Alanine. When a gene is expressed (all the codons in the mRNA are exhausted) each codon in the mRNA produces an amino acid according to the genetic code, and the amino acids are bonded together into a chain.



1	*A*	Ala	Alarine	GCA, GCC, GCG, GCT
2	*C*	Cys	Cysteine (has *S*)	TGC, TGT
3	*D*	Asp	Aspartic acid	GAC, GAT
4	*E*	Glu	Glutamic acid	GAA, GAG
5	*F*	Phe	Phenylalanine[1]	TTC, TTT
6	*G*	Gly	Glycine	GGA, GGC, GGG, GGT
7	*H*	His	Histidine[2]	CAC, CAT
8	*I*	Ile	Isoleucine[3]	ATA, ATC, ATT
9	*K*	Lys	Lysine[4]	AAA, AAG
10	*L*	Leu	Leucine[5]	TTA, TTG, CTA, CTC, CTG, T
11	*M*	Met	Methionine[6] (has *S*	ATG
12	*N*	Asn	Asparagine	AAC, AAT
13	*P*	Pro	Proline	CCA, CCC, CCG, CCT
14	*Q*	Gln	Glutamine	CAA, CAG
15	*R*	Arg	Arginine[7]	AGA, AGG, CGA, CGC, GG, GT
16	*S*	Ser	Serine	AGC, AGT, TCA, TCC, TCG, TCT
17	*T*	Thr	Threonine[8]	ACA, ACC, ACG, ACT
18	*V*	Val	Valine[9]	GTA, GTC, GTG, GTT
19	*W*	Trp	Tryptophan[10]	TGG
20	*Y*	Tyr	Tyrosine[11]	TAC, TAT

Figure (7): *List of the twenty amino acids*

Figure 8 shows an example of how mRNA is converted to protein using the genetic code. When all codons in the mRNA are exhausted, we get a long chain of amino acids (typically a few hundred long). This is the protein corresponding to the original gene. Notice that there is a start codon ATG which signifies the beginning of the protein-coding part. If a start codon occurs inside a gene again, it produces the amino acid methionine. There are three stop codons, which terminate the coding part of the gene. The translation of the codons into amino acids is made by adaptor molecules called transfer RNA (tRNA) molecules. There are more than 20 kinds of tRNA in the cell, at least, one for each amino acid. One end of the molecule matches a specific codon and the other end attaches to the corresponding amino acid. The molecule ribosome works in conjunction with tRNA molecules and mRNA to produce the protein. It is clear that the genetic code is essentially stored in the tRNA molecules. It is a wonder of nature that all life forms (from bacteria to mammals) use the same genetic code. This is no doubt due to the common origin of all life.

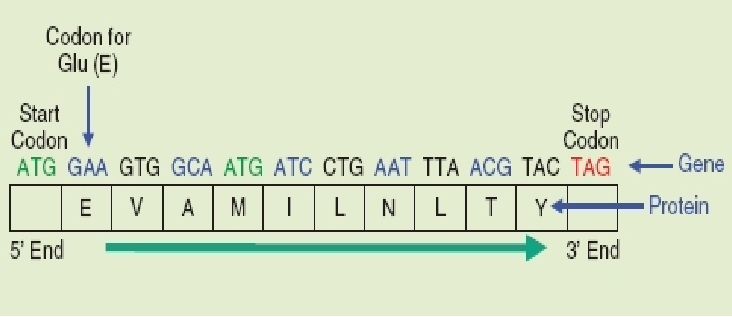


Figure (8): *An example of how mRNA is converted to protein*

Similar to DNA, a protein molecule can be represented by a string of characters from an alphabet of size 20. Because of the innumerable combinations fro the alphabet of 20 amino acids, the number of different proteins in living organisms is enormous. Proteins drive most of the biological processes in living organisms.

3. Methods for identifying protein-coding regions:

In this section, we discuss the application of signal processing in genomics and review existing techniques used for gene identification. There have been various signal processing techniques that have been applied in the field of genomics, due to the discrete nature of the DNA and protein data.

3.1. Numerical mapping techniques:

As these DNA and protein sequences are symbolic, they cannot be processed directly with the existing signal processing algorithms, and numerical assignments need to be made to these sequences. After these symbolic sequences are mapped to suitable numerical sequences, we can apply existing signal processing algorithms to them in order to explain a few of the important properties of the sequences [1, 2].

There are three categories of numerical mapping techniques; indicator sequence mapping, real number mapping, and complex number mapping [6. 9-12]. In indicator sequence mapping, numerical domain mapping for the DNA sequence can be obtained in terms of indicator vectors. In this category, each nucleotide is represented using an indicator vector u_k, wherein only the position of the element k (k = A, T, C, and G) is represented by the number one and all others positions are zero. Then, some weights are assigned to the nucleotides. A binary indicator sequence is obtained by setting the corresponding weight to 1 and the other weights to 0. Consider the sequence $x(n) = \{\text{AATTCAGGCTAGTCTAACC}\}$. For this sequence, we have the binary indicator sequences as

$$b_A(n) = \{1100010000100001100\},$$
$$b_C(n) = \{0000100010000100011\},$$
$$b_G(n) = \{0000001100010000000\},$$
$$b_T(n) = \{001100000100101000\}. \qquad (1)$$

The indicator sequences composed of ones and zeros. The position corresponding to the presence of the nucleotide is denoted by one, in the respective indicator sequences. For example, $b_A(n)$ has one at positions n = 1,2, 6,11,16,17 since the sequence $x(n)$ has A in the corresponding positions. In real number mapping , the number mapping is as follows

$$A \to -1.5,\ T \to -0.5, C \to 0.5, G \to 1.5\ . \qquad (2)$$

The notation in Equation (2), e.g. A → -1.5, can be read as the nucleotide A is mapped to number -1.5. The complementary nucleotides are equal in magnitude and opposite in sign. This rule is better suited for computing correlation values. Another method that is discussed in [6, 8, 11] assigns an increasing sequence of integers to the alphabetically sorted nucleotides after obtaining the indicator sequences. The assignment is done as

$$A \to 1,\ C \to 2, G \to 3, T \to 4\ . \qquad (3)$$

In complex number mapping, the four characters (A, T, C, and G) are represented by four complex numbers as follows (see [7] for details):

$$A \to 1 + j,\ C \to -1 - j, G \to -1 + j, T \to 1 - j\ . \qquad (4)$$

After these symbolic sequences are mapped to suitable numerical sequences, we can apply existing signal processing algorithms to them in order to explain a few of the important properties of the sequences. Signal processing techniques, such as Fourier transform and wavelet transform, have been used to identify a periodicity in the DNA sequences and to help in finding the protein coding regions of the sequence. This is also called gene identification and is characterized from the frequency spectrum of the DNA sequences. We focus on Fourier transform and digital filtering. A detailed discussion on applying Fourier analysis to DNA sequences is presented in [9].

3.2. Applying Fourier analysis:

The Fourier spectra of the DNA sequences using a sliding window identify the coding regions present in the DNA sequence. It is noticed that protein-coding regions (exons) in genes have a period-3 component because of coding biases in the translation of codons into amino acids [10-13]. The period-3 property is not present outside exons, and can be exploited to locate exons. The major signal that characterizes the coding regions is the three base periodicity. The value of the spectrum at $f = N/3$ (where N is the length of the window) determines the nature of the DNA region. The three-base periodicity is also called the $N/3$ periodicity or $2\pi/3$ periodicity. The spectra of the binary indicator sequences can be computed to obtain the power spectrum as in [9]. A coding measure can be characterized based on the relative strength of the periodicity at $f = N/3$ and appears as a peak in the average spectrum. The origin of the periodicity can be attributed to the codon bias that refers to the unequal usage of codons in the coding regions and the triplet bias, which is the bias in the usage of nucleotide triplets. The advantages of these techniques are that they are robust to sequencing errors and computational complexity is very low. A mathematical treatment is presented in [12] to explain the reason for the peak at $f = N/3$ in a coding region.

Consider a DNA sequence, $D(i), 0 \le i \le N-1,$ consists of four nucleotides, A, T, C, and G. The DNA sequence is mapped into four indicator sequences, $A(i), T(i), C(i),$ and $G(i), 0 \le i \le N-1.$ These soft sequences are related to the presence or absence of the four nucleotides at location i in $D(i)$. For a given i, the value of an indicator sequence, attributed to certain nucleotide, takes a value of 1 at location i if the i^{th} element of the DNA sequence declare the same type of nucleotide. Else, the value of the soft function takes a value zero.

For example, for the DNA sequence $(N = 24)$,

$$D(i)=\{C\ T\ G\ C\ A\ T\ G\ A\ C\ T\ A\ A\ G\ A\ G\ T\ C\ C\ G\ T\ A\ T\ G\ A\}, \qquad (5)$$

the four indicator sequences are

$$A(i)=[0,0,0,0, 1,0,0,1,0,0,1,1, 0, 1,0,0, 0,0,0,0, 1,0,0,1],$$

$$T(i)=[0,1,0,0, 0,1,0,0,0,1,0,0, 0, 0,0,1, 0,0,0,1, 0,1,0,0],$$

$$C(i)=[1,0,0,1, 0,0,0,0,1,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0],$$

$$G(i)=[0,0,1,0,0,0,1,0,0,0,0,0,1,0,1,0,0,0,1,0,0,0,1,0]. \qquad (6)$$

The four indicator sequences are parsed into overlapping segments of length $L,\ L \le N$. This can be achieved by using a window w of length L and slide the window forward one by one until all the soft sequences are processed. In most publications [9-11, 13-15], a rectangular window is used. The following rectangular window is used to parse the indicator sequences:

$$w(n) = \begin{cases} 1, & 0 \le n \le (L-1), \\ 0, & \text{otherwise.} \end{cases} \qquad (7)$$

For the truncated soft sequences, the discrete Fourier transform (DFT) is calculated as

$$U_B(k) = \sum_{n=0}^{L-1} B(n)\ w(n) e^{(-j2\pi k n/L)}, B \in (A,T,C,G), k = 0,1,2,......., L-1. \qquad (8)$$

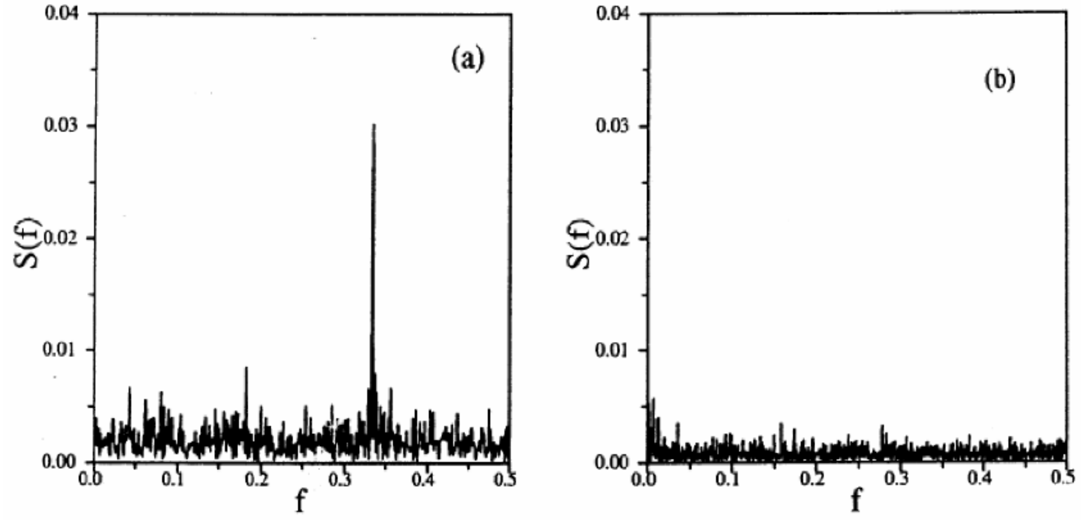
The sum of the absolute values of the power spectra is determined as

$$S(k) = \sum_{n=0}^{L-1} (|U_A(k)|^2 + |U_T(k)|^2 + |U_C(k)|^2 + |U_G(k)|^2),\ \ k = 0,1,2,......., L-1. \qquad (9)$$

The sum of the absolute values of the power spectra, $S(k)$, is used as indicator of coding regions. It is used as a coding measure to detect probable protein-coding genes in the DNA sequence. When $S(k)$ is plotted against k, it reveals an appreciable peak at coding region ($N/3$) and shows no such peak for non-coding region. Fig. 9-a shows the absolute value of the power spectral density (PSD) in case of a single exon coding region, while Fig. 9-b shows the case of non-coding regions. Figure 10 shows a coding region inside a genome of baker's yeast (N=1320). Figure 11 shows the case of two protein-coding regions and indicates the true axon locations. Figure 12 shows the exon prediction results for gene F56F11.4 in the C elegance chromosome III. It shows five exons.

Most of the methods presented in the literature (see [1, 5] for examples) use binary indicator sequences and rectangular windows to parse the DNA sequence under processing. The binary sequences and rectangular windows cause abrupt truncations of the DNA sequence and leads to extraneous peaks in the spectrum. Some other methods use soft sequences and apply a gradual window to parse the DNA sequence. These methods avoid abrupt truncations of the DNA sequence and remove most of the extraneous peaks which improve the discrimination results.

An extension to the Fourier transform approach has been proposed in [1, 7], where DNA spectrograms (squared magnitude of the short-time Fourier transform) are computed in order to differentiate between the coding and non-coding regions. In order to compute the spectra, two numerical mapping techniques, binary mapping and complex mapping, are used. The DNA spectrum is computed using the short-time Fourier transforms by translating windows along the binary indicator sequences. The weighted spectrum is then computed by reducing the dimensionality of the indicator sequences from four to three. The weights are optimized such that the observed peaks at $f = N/3$ can be distinguished from the non-coding regions, i.e., the other periodicities.



(a) Case of a single exon coding region *(b) Case of a non-coding regions*

***Figure (9):** Power spectral density of coding* and non-coding regions

3.3. Using digital filtering:

The binary indicator sequences with sliding window method can be regarded as digital filtering followed by downsampling. The rate of downsampling depends on the separation between adjacent positions of the window. The corresponding digital filter has an impulse response

$$h(n) = \begin{cases} e^{jw_0 n} & \leq 0 n \leq L-1 \\ 0 & \text{otherwise.} \end{cases} \qquad (10)$$

This digital filter is a bandpass filter with passband centered at ω0=2π/3 and minimum stopband attenuation of about 13 dB (Figure 13). If we take the design and implementation of digital filters into consideration, we can isolate the period-3 behavior. There are efficient methods to design and implement the filters to be suited to gene prediction application [15]. Assume a narrow bandpass digital filter $H(z)$ with passband centered at ω0=2π/3 with an input indicator sequence $x_G(n)$, where n is the base location, and an output sequence $y_G(n)$. In the protein-coding region, the input sequence $x_G(n)$ is expected to achieve the period-3 property. If this is true, i.e. if the input sequence $x_G(n)$ have a period-3 component, then it has large energy in the digital filter passband. In this case, the output sequence $y_G(n)$ will be relatively large in the coding regions. Fig. 14 shows this scenario in case when the input sequence $x_G(n)$ having a coding region. Similarly we can define a sequence related to the summation of the four bases sequences $\{A(i), T(i), C(i), G(i)\}$ as

$$Y(n) = |y_A(n)|^2 + |y_T(n)|^2 + |y_C(n)|^2 + |y_G(n)|^2 \,. \qquad (11)$$

The plot of $Y(n)$ can be used as an indication of the coding regions.

Figure 15 shows the exon prediction results for gene F56F11.4 in the C. elegance chromosome III. This gene has five exons. The first plot (left plot) uses the DFT based spectrum using a sliding window. The five peaks corresponding to the exons are shown in Fig. 15. The second plot (right plot) uses a multistage filter H(z) [16]. Note that the five exons are seen very clearly in this case. Figures 16-a and -b show the same comparison in case of the genome HUMCBRG. It is clear from Figures 15 and 16 that using digital filtering achieves better discrimination results. Further digital filtering design details can be found in [16]. 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 code a given amino acid, they are not used with uniform probability in organisms. For example, base G dominates at certain codon positions in the coding regions. It is observed that the use of the plot $y_G(n)$, which depends on base G alone, is sufficient for revealing the period-3 property, and therefore for the prediction of protein coding regions. The use of digital infinite-impulse response (IIR) filtering to detect coding regions in the DNA sequences has been presented in [3]. As the $N/3$ periodicity is exhibited by the coding region in a DNA sequence, an anti-notch filter to identify these coding regions is discussed in [3]. The design of the anti-notch filter is based on the fact that there is a sharp peak at 2π/3 in the spectrum. Again, this is a bandpass filter with passband centered at $\omega_0 = 2\pi / 3$.

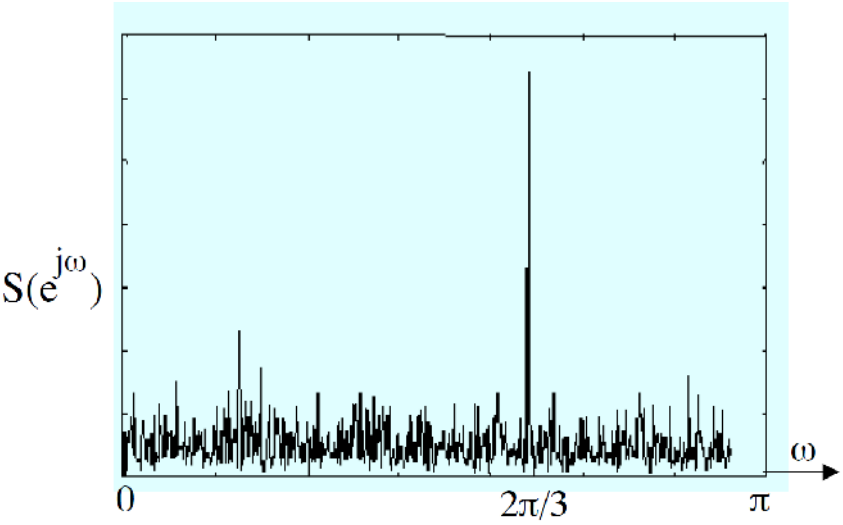


Fig. 10 Coding region of length N=1320 inside a genome of baker's yeast

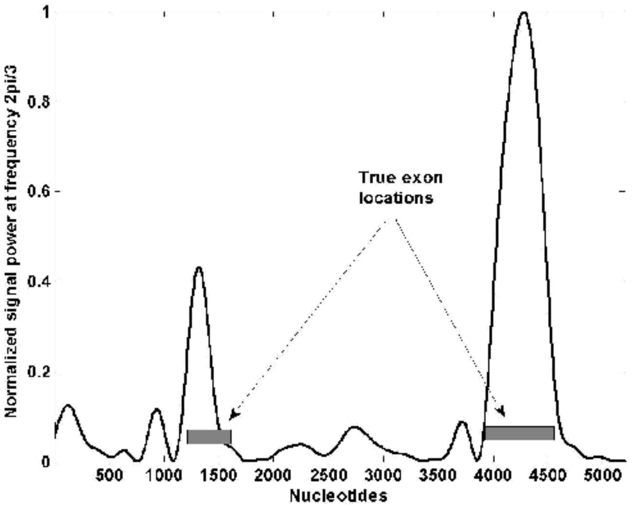


Fig.11 Normalized signal power in case of two exons

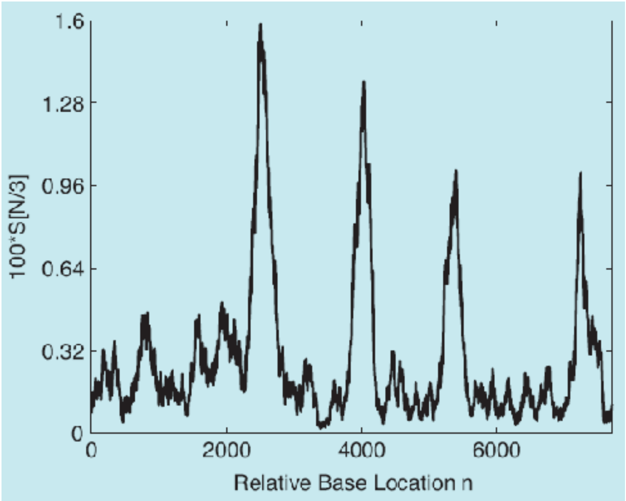


Fig. 12 The DFT spectrum S(N/3) for gene F56F11.4 in the C. elegance chromosome III

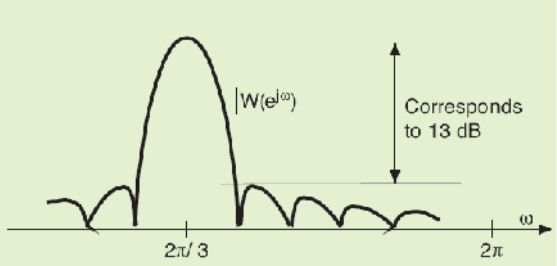


Fig. 13 A digital filter with passband centered at $\omega_0 = 2\pi/3$

Fig. 14 A digital filter with passband center frequency $\omega_0 = 2\pi/3$, *input sequence* $x_G(n)$, *and output sequence* $y_G(n)$

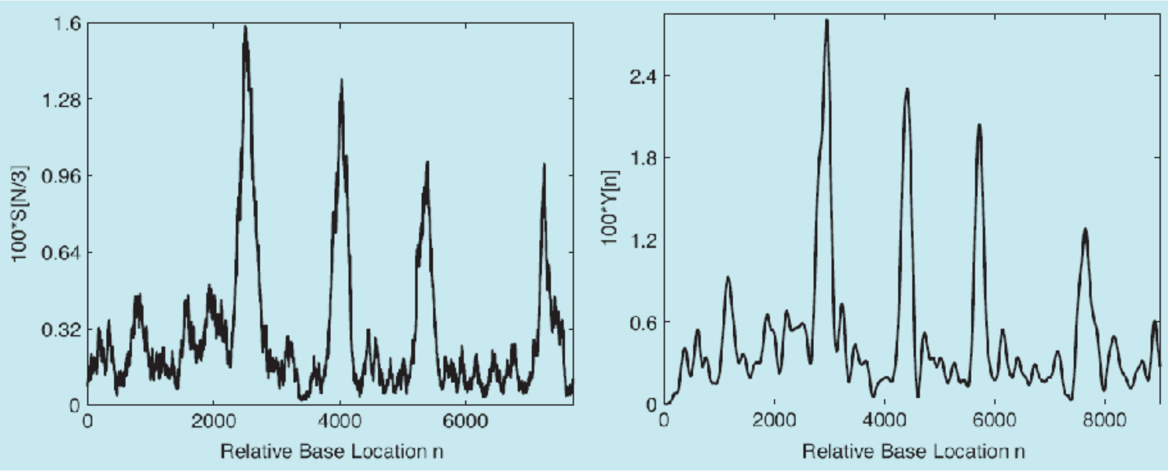
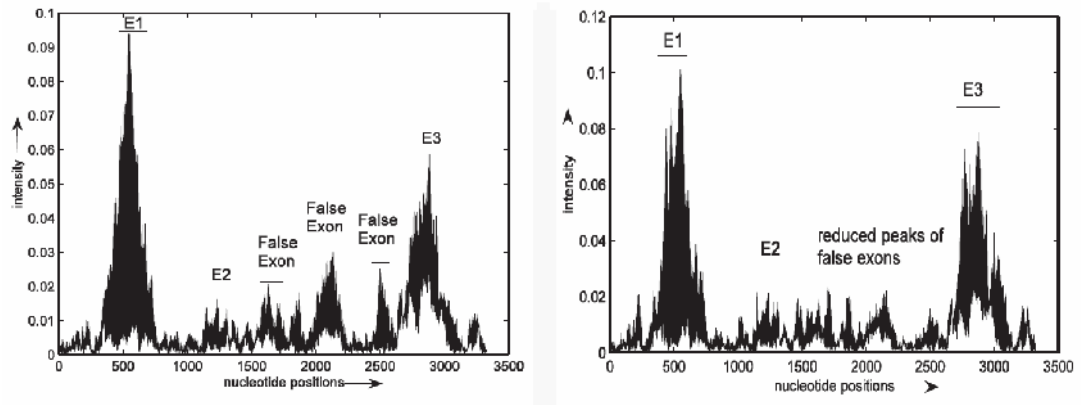


Fig. 15 Left plot: The DFT spectrum S(N/3) for gene F56F11.A in the C. elegance chromosome III; Right plot: the bandpass filter output for the same gene.



(a) Using power spectral density *(b) Using band pass filter*

Fig. 16 Spectrum of HUMCBRG using power spectral density and band pass filter

4. Conclusions:

In this paper, the molecular biology of DNA has been discussed from the signal processing point of view. An outline of the basic biology and structure of DNA have been discussed. A brief review of the applications of signal processing theory in the field of bioinformatics has been presented. Most of the existing methods depend on the pronounced period three peaks observed in the Fourier transform of the coding regions in genes using binary indicator sequences and rectangular window. Some other methods rely on soft indicator sequences and gradual windows. It has been shown that these methods reveal period three peaks for coding regions and show no such peak for non-coding regions. It has been shown that the methods with digital filtering achieve better discrimination between the coding areas and non-coding regions, based on experimental data of a number of genes, compared to the methods using Fourier transform.

References:

[1] P. P. Vaidyanathan, *Genomics and Proteomics: A Signal Processor's* Tour, IEEE Circuits and Systems Magazine, pp. 6-29, 2004.

[2] Ré Matteo and Pavesi Giulio, *Detecting Conserved Coding Genomic Regions through Signal Processing of Nucleotide Substitution Patterns*, Artificial Intelligence in Medicine, Vol. 45, No. 2-3, pp.117-123, March 2009.

[3] J. D. Watson and F. H. C. Crick, *"A Structure for DNA,"* Nature, p. 737, April 1953.

[4] Z. Aydin and Y. Altunbasak, *A Signal Processing Application in Genomic Research: Protein Secondary Structure Prediction,* IEEE Signal Processing Magazine, Vol. 7, pp.128-131, July 2006.

[5] D. Schonfeld, J. Goutsias, I. Shmulevich, I Tabus, and A. H. Tewfik, *Introduction to the Issue on Genomic and Proteomic Signal Processing,* IEEE Journal of Selected Topics in Signal Processing, Vol. 2, No. 3, pp.257-260, June 2008.

[6] D. Anastassiou, *Genomic Signal Processing,* IEEE Signal Processing Magazine, pp. 8-20, July 2001.

[7] J. Berger, S. Mitra, and J. Astola, *Power Spectrum Analysis for DNA Sequences,* IEEE Transactions on Signal Processing, Vol. 2, pp. 29-32, 2003.

[8] Jaakko Astola, Edward Dougherty, Ilya Shmulevich, Ioan Tabus, *Genomic Signal Processing,* Signal Processing, Vol. 83, No. 4, pp.691-694, April 2003.

[9] M. K. Hota and V. K. Srivastava, *DSP Technique for Gene and Exon Prediction Taking Complex Indicator Sequence,* TENCON 2008, IEEE Region 10 Conference, pp.1-6, Nov. 2008.

[10] Ilya Shmulevich, *Genomic Signal Processing,* Princeton Univ. Press., New York, 2007.

[11] M. Akhtar, J. Epps, and E. Ambikairajah, *Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction,* IEEE Journal of Selected Topics in Signal Processing, Vol. 2, No. 3, pp.310-321, June 2008.

[12] S. Datta and A. Asif, *A Fast DFT Based Gene Prediction Algorithm for Identification of Protein Coding Regions*, Proc. IEEE International Conference on Acoustic, Speech, and Signal Processing, ICASSP, Vol. 5, pp.653-656, March 2005.

[13] D. Anastassiou, *Frequency Domain Analysis of Biomolecular Sequences,* Bioinformatics, Vol. 16, No. 12, pp. 1073-1081, 2000.

[14] Paul Dan Cristea, *Large Scale Features in DNA Genomic Signals,* Signal Processing, Vol. 83, No. 4, pp.871-888, April 2003.

[15] P. P. Vaidyanathan and B-J. Yoon, *Digital Filters for Gene Prediction Applications*, IEEE Asilomar Conference on Signals, Systems, and Computers, Monterey CA, pp. 871:879, Nov. 2002.

[16] Y. Neuvo, C. Y. Dong, and S. K. Mitra, *Interpolated Finite Impulse Response Filters,* IEEE Transactions on ASSP, pp. 563-570, June 1984.