GE-INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH
VOLUME -2, ISSUE -8 (October 2014) IF-3.022 ISSN: (2321-1717)

SIGNAL PROCESSING IN PROTEIN CODING DNA FOR CANCER DETECTION

Sujit Kumar Chakravarty,
Electronics & Telecommunication Engineering
BijuPatnaik University of Technology, Odisha, India

ABSTRACT

Protein coding regions in DNA sequences is a fundamental step in computational recognition of genes. An important emerging research area is the study and development of signal processing techniques for rapid real time nucleic acid detection. Signal Processing deals with life forms, particularly the DNA sequences. Such techniques become extremely important in obtaining useful information from the large sets of data in the form of the human genome. This papers deals with the protein-coding regions of the DNA sequence .

Key Words- DNA, Digital Signal, DFT, Filters,

I. INTRODUCTION

Signal processing is the art of representing, transforming, analyzing, and manipulating signals. Genomic engineering is a quickly evolving interdisciplinary field that blends bioscience, medicine, and engineering. DNA sequence analysis technology has been developing for decades to unravel the structure-related information. Signal processing techniques have been found useful in truly diverse applications, such as signal enhancement, speech recognition , audio and image compression , radar signal processing , and digital communications , just to name a few. More recently, signal processing techniques have been also applied to the analysis of biological data with considerable success. A key concept in DSP is the possibility of representing the signals in the frequency domain making use of the Discrete Fourier Transform . This representation leads to some important signal properties which are associated to their frequency spectrum that are not revealed in the time domain. In case of the genomic sequences the nucleotide bases are represented mathematically by character strings of size 4 alphabet consisting of the letters A,T,G and C. The possibility of finding wide applications of DSP techniques to the analysis of genomic sequences occur only when these are appropriately converted into numerical sequences. Now a days Cancer is the most common and dreaded disease that plays a leading role causing death all over the world. Cancer is caused by abnormalities in the genetic material of the transformed cell. Cancer-promoting genetic abnormalities may randomly occur through errors in DNA replication. Nanotechnology is also used to develop accurate and sensitive biomedical devices for cancer genome study. Abnormality of the DNA and coding regions are related to cancer.

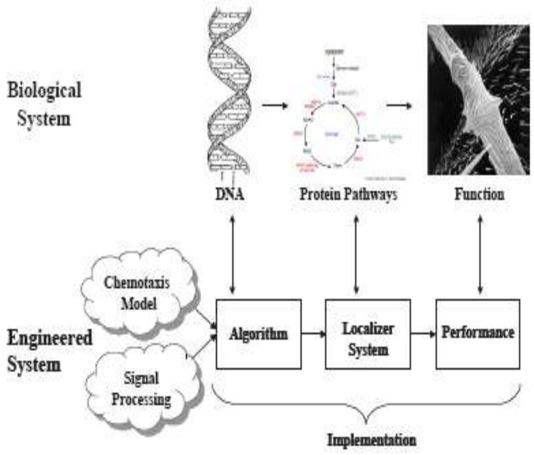


Figure 1. Paralleling an engineered system to a biological system.

II. DNA

A single strand of Deoxyribo nucleic acid (DNA) consists of many linked, smaller components called nucleotides. Each nucleotide is one of four possible aminoacids namely Adenine (A), Thyamine (T), Cytosine (C) and Guanine (G). These are represented by the alphabets A, T, C, and G. DNA has two distinct ends, the 5'end and the 3' end. The 5'end of a nucleotide is linked to the 3'end of another nucleotide by a strong chemical bond, thus forming a long, one dimensional chain of a specific directionality. Single DNA strands tend to form double helices with other single DNA strands. A DNA double strand contains two single strands that are complementary to each other i. e A is linked to T and vice versa, and C is linked to G and vice versa. Each such bond is weak but together all these bonds create a stable, double helical structure . The two strands run in opposite direction. The diagram is a simplified, straightened out depiction of the two linked strands. For example, the part of the DNA double strand shown in Fig. is

5′ - C-A-T-T-G-C-C-A-G-T - 3′
3′ - G-T-A-A-C-G-G-T-C-A - 5′

Because each of the strands of a DNA double strand uniquely determines the other strand, a double-stranded DNA molecule is represented by either of the two character strings read in its 5′ to 3′ direction.

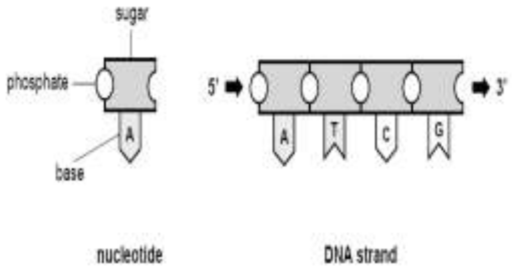


Figure 2: Illustration of a nucleotide and a DNA strand.

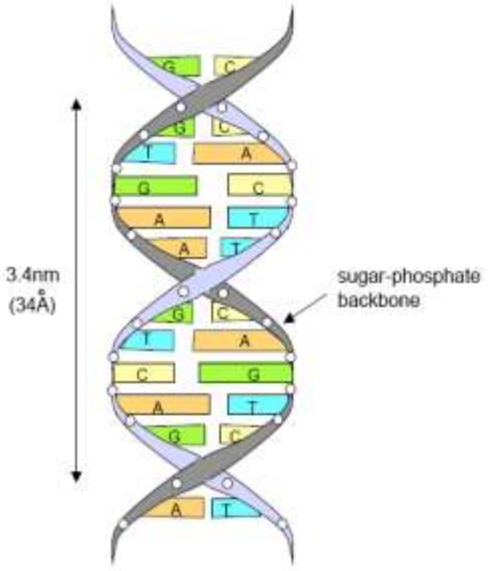
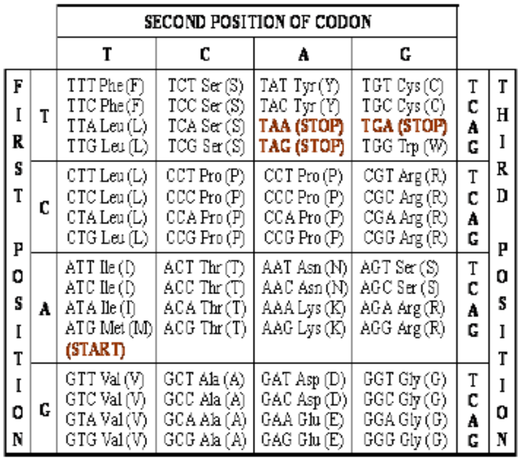


Figure 3: Illustration of a DNA double helix.

III. GENETIC CODE

A protein is a complex molecule consisting of many linked, smaller components called amino acids. Protein synthesis is governed by the genetic code which maps each of the 64 possible triplets (codons) of DNA characters into one of the 20 possible amino acids. The genetic code in which the 20 amino acids are designated by both their one-letter and three-letter symbols. A particular triplet, ATG, serves as the START codon and it also codes for the M amino acid (methionine); thus, methionine appears as the first amino acid of proteins, but it may also appear in other locations. We also see that there are three STOP codons indicating termination of amino acid chain synthesis, and the last amino acid is the one generated by the codon preceding the STOP codon. Coding of nucleotide triplets into amino acids can happen in either the forward or the reverse direction based on the complementary DNA strand. Therefore, there are six possible reading frames for protein coding DNA regions.



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

Figure 4. Genetic code

The total number of nucleotides in the protein coding area of a gene will be a multiple of three, that the area will be bounded by a START codon and a STOP codon, and that there will be no other STOP codon in the coding reading frame in between. However, given a long nucleotide sequence, it is very difficult to accurately designate where the genes are. Accurate gene prediction becomes further complicated by the fact that, in advanced organisms, protein coding regions in DNA are typically separated into several isolated sub regions called exons. When DNA is copied into mRNA during transcription, the introns are eliminated by a process called splicing. The same gene can code for different proteins. This happens by joining the exons of a gene in different ways.

IV. SIGNAL PROCESSING IN DNA SEQUENCE

The DNA sequence contains the instructions that control nearly everything about how an organism lives, such as its development, metabolism, and sensitivity to infection. Its analysis is an important research project in genomic signal processing. Signal processing will play an important role in reaching this goal, and indeed many computational techniques have already been applied, including the artificial neural network , nonlinear model , spectrogram , and statistical techniques . The analysis of DNA sequences using DSP can be useful in the detection of protein coding regions in genomic sequences. In a eukaryotic genome, the introns and exons, start codon and stop codon, donor splice sites (transition from an exon to an intron or vice versa), and a CpG island (a region rich in CG pairs that may promote gene function) can be detected using DSP techniques.

DNA Spectrum detector of Protein Coding

A. Discrete Fourier transform (DFT)

The signals represented in time domain, at times is unable to infer the hidden information and patterns in the signal. Therefore, it is necessary to represent the signal in some alternate domains where the internal characteristics of the signal can be reflected in a better way. The Fourier transform (FT) provides such a representation by transforming a signal from time domain into frequency domain. The Fourier transform is an invertible integral transform that expresses a function in terms of sinusoidal basis functions, i.e. as a sum or integral of sinusoidal functions of different frequencies .

The Fourier transform X(f) of a signal x(t) is defined as

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft\ dt}$$

and its inverse relationship is given by

$$x(t) = \int_{-\infty}^{\infty} X(f).e^{j2\pi ft\ dt}$$

The discrete version of the Fourier transform is called the discrete Fourier transform (DFT). This is used when both the time and the frequency variables are discrete. The DFT of a discrete time signal x(n) of length N can be viewed as a uniformly sampled version of X(f) at frequencies

$f_k = \frac{K}{N}$ for k = 0, 1,………… ,N −1.

The period of the signal is $\frac{N}{K}$ The DFT of the signal x(n) is defined as

$$X\left(\frac{K}{N}\right) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)e^{\frac{-j2\pi nk}{N}}$$

Hence the inverse DFT (IDFT) is defined as

$$x(n)\sum_{k=0}^{N-1} X\left(\frac{K}{N}\right)e^{\frac{j2\pi nk}{N}}$$

The discrete Fourier transform is one of the most common spectral analysis technique and has been used in various fields such as image analysis, filtering, pattern analysis, feature extraction in various areas in engineering.

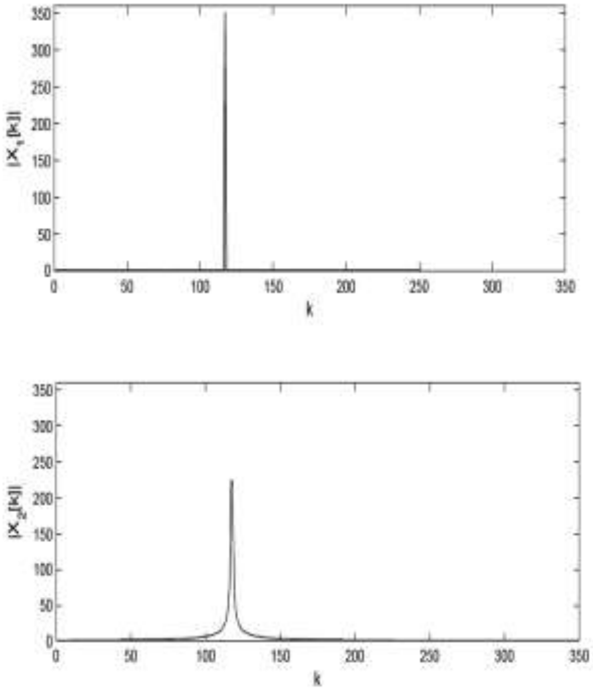


Figure 5: DFTs of periodic signals. (Top) Magnitude plot of the DFT of a signal with a period T = 3. (Bottom)

B. *Digital Filters*

The filter itself has a very simple impulse response

$$\omega(n) = \begin{Bmatrix} e^{j\omega_0 n} & 0 \le n \ge N-1 \\ 0 & otherwise \end{Bmatrix}$$

This is a band pass filter with pass band centered at $\omega 0 = 2\pi/3$ and minimum stop band attenuation of about 13 dB.

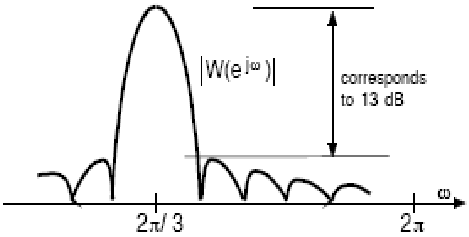


Figure 6. The filtering effect of DFT computation.

A narrow band band pass digital filter *H*(*z*) with pass band centered at $\omega 0 = 2\pi/3$. With the indicator sequence *xG*(*n*) taken as input, let *yG*(*n*) denote its output. The narrow band filter *H*(*z*) can be regarded as an **antinotch filter** (i.e., complement of a notch). A digital filter is a discrete system capable of realizing some transformation to an input discrete numerical sequence. IIR antinotch Filter is a 2nd order all pass filter defined as

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

A filter bank with 2filter G(Z) and H(Z)

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

Then $G(Z) = K\left[\frac{1-2cos\,\omega_0 Z^{-1}+Z^{-2}}{1-2Rcos\,\theta_{z^{-1}}+R^2Z^{-2}}\right]$

Where $cos\omega_0 = \frac{2Rcos\theta}{1+R^2}, K = \frac{1+R^2}{2}$
R is less then and closed to unity G(Z) is a notch filter with a 0 at frequency ω_0. Also H(z) and G(z) are power complementary. Hence H(z) can be a good anti-notch filter

$$H(Z) = \frac{1}{2}\left[\frac{(1-R^2)(1-Z^{-2})}{1-2Rcos\theta_{z^{-1}} + R^2 Z^{-2}}\right]$$

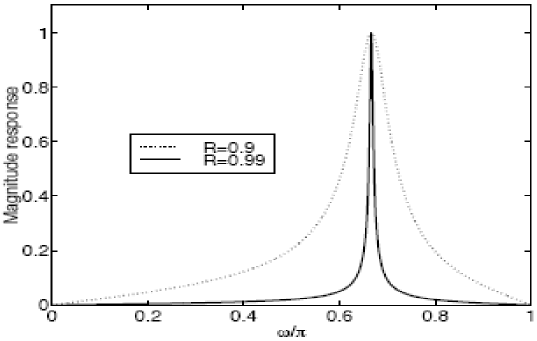


Figure 7. Antinotch filter responses for two values of *R*.

V. RESULT

In order to demonstrate the proposed idea, first took a segment of DNA sequence in chromosome III and computed the four indicator sequences xA(n), xC(n), xG(n), and xT (n). This DNA segment contains the protein-coding gene that consists of five exons. Firstly, used the DFT-based method to compute S[N/3]. Figure 8 (Top) shows the plot of S[N/3] as function of relative base location (n = 0). As shows in Figure 8 (Top), the last four exons have clearly visible peaks. However, the peak that arises from the first exon is somewhat buried in the noisy background and it is not easily distinguishable from the spurious peaks. The plot in Figure 8 (Bottom) shows the output Y (n) obtained by the allpass-based antinotch filter with a pole radius R = 0.992. In this plot, the first peak is also larger than the spurious peaks in the background, showing the location of the first exon more clearly. This result shows that the antinotch filter approach works very well, while providing additional advantages in implementation.

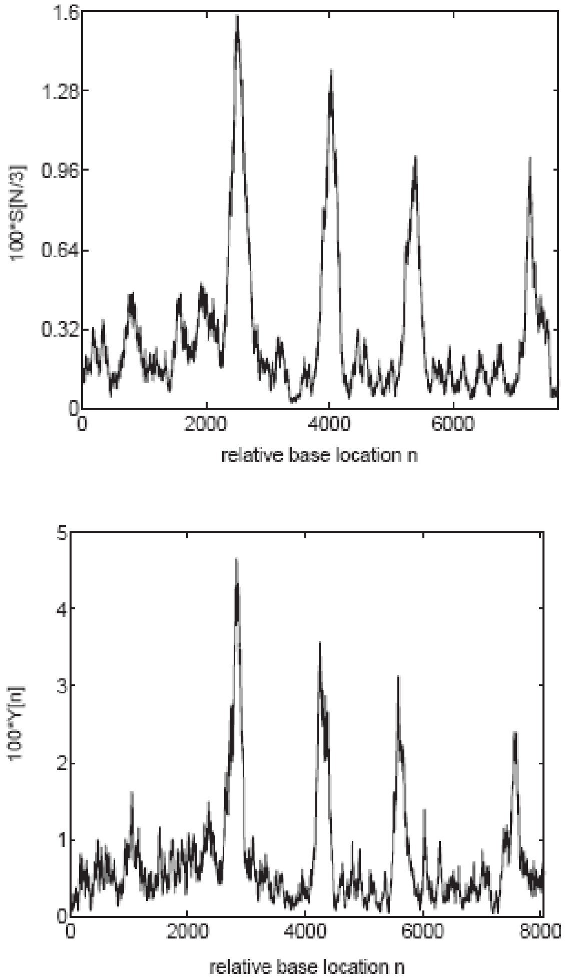


Figure 8 Exon prediction results for chromosome III. (Top) Plot of S[N/3] computed using the DFT. (Bottom) Plot of Y (n) that is computed using the antinotch

VI. CONCLUSION

DSP now-a-days plays an important role in DNA sequence analysis, CANCER diagnosis and gene expression analysis etc. Researchers are using DFT power spectrum plot to predict protein coding regions of a DNA sequence. DFT power spectrum as a method to predict CANCER disease for various databases available in Gene bank. The filtered power spectrum plots yield high accuracy. The paper have conducted some preliminary studies about the prediction of CANCER cells . Further efforts will be made to improve the accuracy of prediction by using other types of digital filters. Signal processing-based computational and visual tools are meant to synergistically complement character-string-domain tools that have successfully been used for many years by computer scientists. In this paper, some of several possible ways that signal processing can be used to directly address biomolecular sequences are illustrated. An important advantage of DSP-based tools is their flexibility. parameters in ways that will enhance the appearance of these patterns, thus clarifying their significance.

VII. REFERENCE

[1] P. P. Vaidyanathan and B. Yoon, “ The role of Signal-Processing Concepts in Genomics and proteomics”, J. Franklin Inst., vol. 341, pp. 111-135, 2004.

[2] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, Essential Cell Biology: An Introducton to the Molecular Biology of the Cell, New York, NY: Garland Publishing Inc., 1997.

[3] D. Anastassiou, “Genomic signal processing,” IEEE Signal Processing Magazine, vol. 18, pp. 8-20, 2001.

[4] A. Bairoch and P. Bucher, “PROSITE: Recent developments,” Nucleic Acids Research, vol. 22, pp. 3583-3589, 1994.

[5] S. Datta and A. Asif, “DFT based DNA splicing algorithms for prediction of protein coding regions,” Proceedings of the 30th IEEE

International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, March 2005.

[6] E. R. Dougherty and A. Datta, "Genomic signal processing: Diagnosis and therapy," IEEE Signal Processing Magazine, vol. 22, pp. 107-112, 2005.

[7] P. Duhamel and M. Vetterli, "Fast Fourier transforms: A tutorial review and a state of the art," Signal Processing, vol. 19, pp. 259-299, 1990.

[8] J. W. Fickett and C. S. Tung, "Assessment of protein coding measures," Nucleic Acids Research, vol. 20, pp. 6441-6450, 1992.

[9] P. D. Cristea, "Conversion of nucleotides sequences into genomic signals," *J. Cell. Mol. Med.*, vol. 6, no. 2, pp. 279- 303, 2002.

[10] D. Anastassiou. DSP in genomics. In *Proceedings of the IEEE International Conference ICASSP 2001*, May 2001.

[11] D. Anastassiou. Genomic signal processing. *IEEE Signal Processing Magazine*, July 2001.

[11] Hayes. M.H:Statistical digital signal processing and modelling. John Wiley & Sons, Inc., New York, USA,1996

[12] E. Sejdic, I. Djurovic and J. Jiang," Timefrequency feature representation using energy concentration:An overview of recent advances", Digital signal processing, 2008.

[13] R. Guan and J. Tuqan, "IIR filter deesign for gene identification," Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS), Baltimore, MD, May 2004.