Study of DNA Sequence Analysis Using DSP Techniques

Inbamalar T M and Sivakumar R
R M K Engineering College, Chennai, India
Email: tminbajustus@gmail.com, rsk.ece@rmkec.ac.in

***Abstract*—Recently there are greater advances in bioinformatics and genomic signal processing. Digital Signal Processing (DSP) applications in genomic sequence analysis have received great attention in recent years. New methods are being developed to analyze Deoxyribonucleic acid (DNA) sequences. In order to use DSP principles to analyze DNA sequences, the DNA sequences should be converted into numeric sequences. Then the DSP algorithms are used in DNA analysis. Discrete Fourier Transform (DFT), digital filtering, Discrete wavelet transform (DWT), Parametric modeling and entreopy are some of the important DSP concepts used for DNA analysis. In this paper, we present a review of the DSP techniques used for DNA analysis. We have explained the required basics behind molecular biology. We have discussed about the applications of DSP in sequence analysis. The DSP techniques used in the literature are studied and their results are compared.**

***Index Terms*—digital signal processing, DNA sequences, exons.**

I. Introduction

DNA sequence analysis has already been a major research topic among computer scientists, physicists, and mathematicians. Digital Signal Processing (DSP) is an important area of science and engineering that has developed as a result of the constant evolution of computer science and technology. DSP comprehends the representation, transformation and manipulation of digital signals as well as the information associated to them. Signals are usually physical magnitudes that vary in time or space. Genomic signals do not have time or space as the independent variable, as occur with most physical signals. Digital signals are those represented as sequences of numbers, as in the case of time series. But genomic sequences can be represented mathematically by character strings of symbols from 4 alphabet sequences consisting of the letters A, T, G and C, which represent each one of the nucleotide bases. In the case of proteins, the alphabet size is 20, corresponding to the possible amino acids. The main reason that the field of signal processing does not yet have significant impact in the field is because it deals with numerical sequences rather than character strings. But, if we properly map a character string into one or more numerical sequences, then digital signal processing (DSP) provides a set of novel and useful tools for solving highly relevant problems [1]-[3].

This paper is organized in the following way. Firstly an overview of the essential concepts from molecular biology is explained, Then the different techniques used for numerical representation of genomic sequences is presented. This allows the application of DSP tools to study genomic sequences. Next, a review of the major applications of DSP to the analysis of genomic sequences is realized, such as identification of protein coding DNA regions, identification of reading frames, location of splice sites and others. We finally review main DSP algorithms used in genomic sequence analysis such as digital filters, the Discrete Fourier Transform (DFT), the Short-Time Fourier Transform (STFT), parametric models (AR, MA, ARMA), Wavelet Transform and the Information Theory concept of entropy.

II. Basic Concepts From Molecular Biology

A. Central Dogma of Molecular Biology

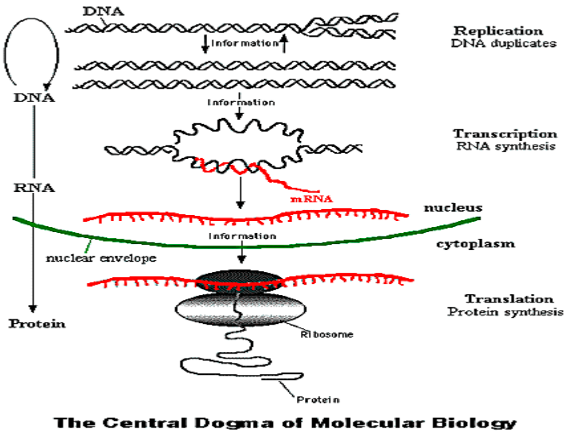


Figure 1. Central dogma of molecular biology.

The Central dogma of molecular biology is that DNA codes for RNA and RNA codes for proteins. Thus the production of a protein is a two-stage process, with RNA playing a key role in both stages. In the first stage, called transcription, a gene within the chromosomal DNA is copied base by base into RNA. The resulting RNA

Manuscript received March 10, 2013; revised May 10, 2013.

doi: 10.12720/joace.1.4.336-342

transcript of the gene is then transported within the cell to a molecular machine called the ribosome that has the task of translating the RNA into a protein. This process is shown in Fig. 1.

B. 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 ie 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 [4] and [5]. The two strands run in opposite directions, as shown in Fig. 2.

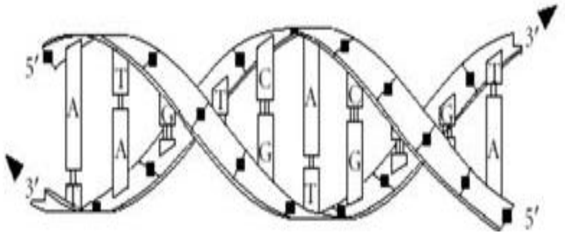


Figure 2. Double helix Structure of DNA

In 1953, James D. Watson and Francis Crick proposed this double helical structure of DNA and they shared the Nobel Prize in 1962 for Physiology and Medicine with Maurice Wilkins who, with Rosalind Franklin, provided the data on which the structure was based.

C. RNA

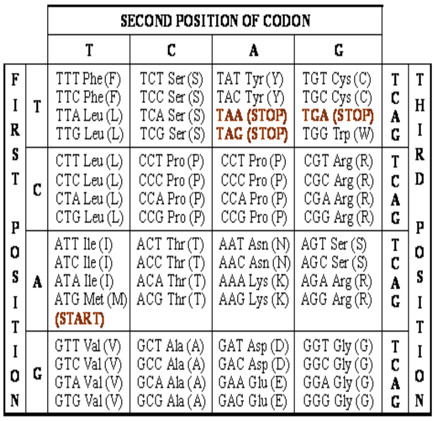
Ribonucleic acid (RNA) is a chemical similar to a single strand of DNA. RNA delivers DNA's genetic message to the cytoplasm of a cell where proteins are made. Ribonucleic acid (RNA) is a ubiquitous family of large biological molecules that perform multiple vital roles in the coding, decoding, regulation and expression of genes. Together with DNA, RNA comprises the nucleic acids, which, along with proteins, constitute the three major macromolecules essential for all known forms of life. Like DNA, RNA is assembled as a chain of nucleotides, but is usually single-stranded. Cellular organisms use messenger RNA (mRNA) to convey genetic information.RNA is often notated using the letters G, A, U, and C for the nucleotides guanine, adenine, Uracil and cytosine that direct synthesis of specific proteins.

D. Proteins

A protein is a complex molecule consisting of many linked, smaller components called amino acids. There are 20 possible types of amino acids in proteins. They are connected with strong bonds, one after the other, forming a long one-dimensional chain (backbone) of a specific directionality. Therefore, as in DNA, a character string mathematically represents each protein. Protein molecules tend to fold into complex three-dimensional (3-D) structures forming weak bonds between their own atoms. They are responsible for carrying out nearly all of the essential functions in the living cell.

E. Genetic Code

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. Fig. 3 shows 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 3. 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. The regions between two successive exons are called introns as shown in Fig. 4. 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. This is called alternative splicing. Alternative splicing seems to be one of the main purposes for which the genes in eucaryotes are split into exons. The mRNA obtained after splicing is uninterrupted and is used for making proteins.

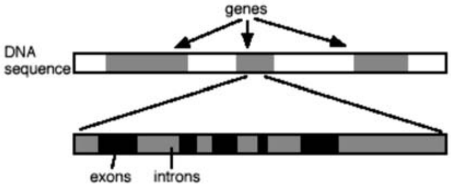


Figure 4. Exons and Introns

F. Gene Regulation

A magical interplay between proteins and DNA is responsible for many of the essential processes inside all living cells. Typically, each gene is being activated or expressed (starting the process that will eventually lead to information processing, and their analysis is one of the most exciting future topics of research that will require a systems-based approach involving cross-disciplinary collaboration at various levels of abstraction, including a genomic level, a macromolecular binding level, and a higher network level.

G. Public Databases

Most of the identified genomic data is publicly available over the Web at various places worldwide, one of which is the entrez search and retrieval system of the National Center for Biotechnology Information NCBI) at the National Institutes of Health (NIH). The NIH nucleotide sequence database is called GenBank and contains all publicly available DNA sequences. There are approximately 126,551,501,141 bases in 135,440,924 sequence records in the traditional GenBank divisions and 191,401,393,188 bases in 62,715,288 sequence records in the WGS division as of April 2011. The complete release notes for the current version of GenBank are available on the NCBI ftp site. A new release is made every two months. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA Databank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis**.** As another example, a specialized repository for the processing and distribution of 3-D, macromolecular structures can be found in the Protein Data Bank at www.rcsb.org.

III. Signal Processing for DNA Sequences

As explained Genomic information is in the form of alphabets A, T, C and G. Signal processing deals with numerical sequences. Hence character strings have to be mapped into one or more numerical sequences. Then signal processing techniques can be applied for analysis of DNA sequences.

A. Numeric Representation

In a DNA sequence we have to assign numbers to the characters *A*, *T*, *C*, *G*, respectively. A proper choice of the numbers *can* provide potentially useful properties to the numerical sequence. For example, if we choose complex conjugate pairs T =A* and *G* = *C** , then the complementary DNA strand is represented conjugate, symmetric numerical sequences which have interesting mathematical properties, including generalized linear phase. Equation (1) shows the complex conjugate pair assignment

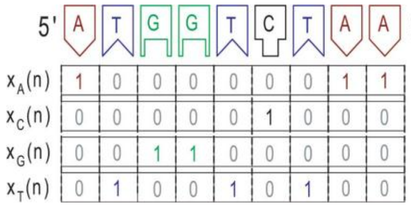
$$A = 1+ j,\ T = 1- j,\ C = -1- j,\ G = -1+ j \qquad (1)$$

Given a DNA sequence, the *indicator sequence* for the base *A* is a binary sequence,

e.g., $x_A(n)$ = 000110111000101010 . . .

where 1 indicates the presence of an *A* and 0 indicates its absence. The indicator sequences for the other bases are defined similarly. This type of representation is called Voss representation [6]. The Voss representation for a specific DNA sequence is shown in Figure 5.

An alternative to the binary sequence method is the electron–ion interaction potential (EIIP) values for nucleotides [7]. Given a DNA sequence, a numerical sequence can be assigned to it such that is equal to the EIIP value of the nucleotide in the DNA sequence. The EIIP values for the nucleotides are given in Table I.



5'	A	T	G	G	T	C	T	A	A	3'
$x_A(n)$	1	0	0	0	0	0	0	1	1	
$x_C(n)$	0	0	0	0	0	1	0	0	0	
$x_G(n)$	0	0	1	1	0	0	0	0	0	
$x_T(n)$	0	1	0	0	1	0	1	0	0	

Figure 5. Voss representation

TABLE I. EIIP Values of Nucleotides

Nucleotide	EIIP
A	0.1260
G	0.0806
T	0.1335
C	0.1340

B. Applications of DSP in the Anlysis of DNA

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. The three-base periodicity was found to be a characteristic of the protein-coding regions in both prokaryotic and eukaryotic sequences. This can be found using DSP techniques. Gene prediction using DSP techniques is another important application. Classification of the DNA sequences can also be done. Reading frame identification is an important issue in the detection of coding regions which can be done with DSP [8-15].

IV. DSP Algorithms for DNA Analysis

A. Discrete Fourier Transform

The Discrete Fourier Transform is a mathematical operation that transforms the time domain function $x[n]$ to frequency domain representation $X[k]$ [16]. The Discrete Fourier Transform evaluates the frequency components required to reconstruct the finite segment of the sequence that was analyzed. The DFT is usually represented in terms of the corresponding magnitude and phase functions that constitute the frequency spectrum of the sequence $x[n]$. The Discrete Fourier transform is a very useful tool, because it can reveal periodicities in the input data as well as the relative intensities of these periodic components. A spectrum obtained by taking DFT for a DNA sequence is shown in Fig. 6.

Using the DFT for spectral analysis of random signals require certain considerations to obtain a statistically valid result. For stationary random signals, a commonly employed procedure to obtain a power spectral density (PSD) function in the frequency domain is the Welch's modified periodograms method. The PSD function is obtained in this case by calculating the mean value of the squared DFT coefficients at each frequency value, for adjacent and usually overlapping windowed signal segments. The measure obtained in this way is a consistent estimate of the power spectrum [17]-[20].

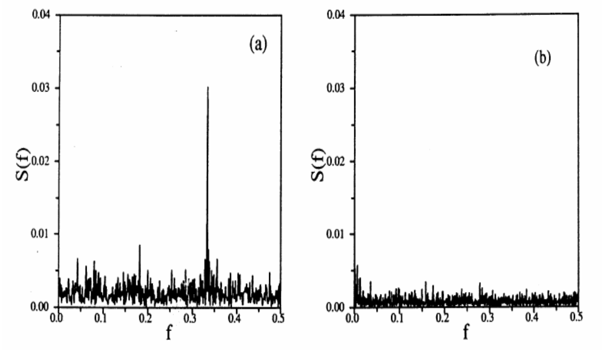


Figure 6. (a)Spectrum of Protein coding region (b) Spectrum of Non coding region

In the case of non-stationary signals, The Short Time Fourier Transform (STFT) is an algorithm frequently used for the DFT-based spectral analysis [21]. In the STFT, the time signal is divided into short segments (usually overlapped) and a DFT is calculated for each one of these segments. A three dimensional graph called *spectrogram* is obtained by plotting the squared magnitude of the DFT coefficients as a function of time.

B. Digital Filters

A digital filter is a discrete system capable of realizing some transformation to an input discrete numerical sequence. There are different classes of digital filters namely linear, nonlinear, time-invariant or adaptive. Digital filters are characterized by numerical algorithms that can be implemented in any class of digital processors. In particular, LTI digital filters can pertain to one of two categories, according to the duration of their response to the impulse, or Dirac delta function, when it is used as the input signal: infinite (IIR) or finite (FIR) impulse response. The system transfer function relates the input and output sequences $x[n]$ and $y[n]$, through their respective Z transforms $X[z]$ and $Y[z]$. A variety of digital filter design techniques allow to obtain any desired magnitude response with frequency selectivity properties. According to the frequency interval transmitted, the magnitude of the basic ideal prototype filter frequency responses can be lowpass, highpass, bandpass and bandstop. A combination of these responses leads to a *multiband* filter. The typical ideal frequency responses (in magnitude) of the prototype filters are shown in Fig. **7**. The window length should be long enough, so that the periodicity effect can be captured above the random variations. But a long window implies more computations and also poorer base domain resolution, while with shorter window the level of noise is increased. Here we have taken the window of lengths 351 & 651 and recalculate the F (N/3) & F (N/3) in each case. The gene used is F56F11.4 in the C-elegans Chromosome III over the base from 7021 to 15120. This 8100 length DNA sequence contains five known coding regions as shown in Table II.

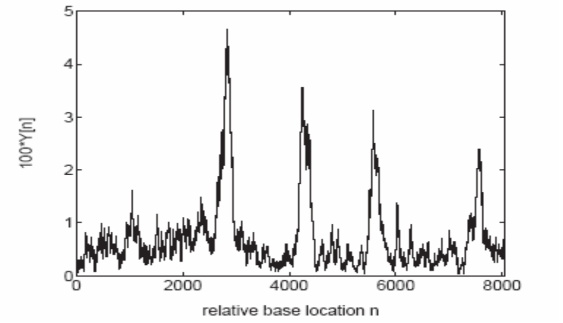


Figure 7. Output of a Bandpass filter showing exons

TABLE II. The Five-Coding Region in Gene F56F11.4

Exon	Relative Location	Length
1	929 - 1135	207
2	2528 - 2857	330
3	4114 4377	264
4	5465 - 5644	180
5	7255 - 7605	351

The Table III below shows SNR values for particular window lengths. With the same window, higher SNR is achieved if the length of window is increased from 351 to 651. IIR anti-notch filters are used for gene prediction. The sharpness of these filters depends on the value of R. Even though the IIR anti-notch method has been found to work well, but with a slight increase in the number of multipliers we can design filters with much better stop band 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. The method to be presented is based on the multistage filtering [22].

TABLE III. SNR VALUES OF VARIOUS WINDOWS AGAINST THEIR LENGTH

Window Name	Using 1^{st} Method		Using 2^{nd} Method	
	Window Length N=351	Window Length N=651	Window Length N=351	Window Length N=651
Rectangular	0.3282	0.3439	0.3111	0.3948
Bartlett	0.4736	0.5483	0.6093	0.9060
Hamming	0.4809	0.5529	0.6301	0.9265
Hanning	0.4989	0.5741	0.6919	0.9993
Blackman	0.5263	0.6284	0.7895	1.1573
Parzen	0.5379	0.6583	0.8374	1.2387
Chebyshev A_s=100	0.5390	0.6626	0.8426	1.2494
Kaiser (=21)	0.5385	0.7126	0.8803	1.3505

C. Parametric Models

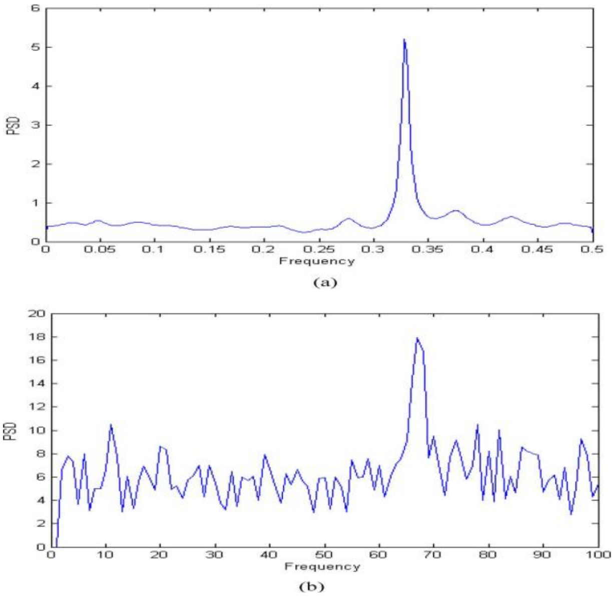


Figure 8. (a) PSD of a sequence X64775 using AR model (b) PSD of same using FT

Parametric spectral analysis is a method that can be used in many cases with some advantages over the non-parametric methods. Its advantages rely in that it is possible to obtain a parametric description of the second-order statistics of a random sequence, by assuming a certain production model for it. A comprehensive analysis of such methods is given in Stoica and Moses. According to the characteristics of the PSD for the analyzed random sequence there are three types of parametric models: Autoregressive (AR) models, corresponding to all pole transfer function, Moving average (MA) models, which correspond to an all-zero transfer function and Autoregressive, moving average (ARMA) models, which is the general case in which there are poles and zeros in the model's transfer function. AR models are more used because of the relative simplicity in calculating the model's parameters through the Yule-Walker equations. The first 200 nucleotides of sequence X64775 are used to compare the FT- and AR-model-based methods. The PSDs from the FT and the AR models are presented in Fig. 8. It is obvious from a comparison of Figure (a) and (b) that the AR model produces a sharp peak and low background noise, while the FT produces a very noisy background. [23]-[25].

D. DWT

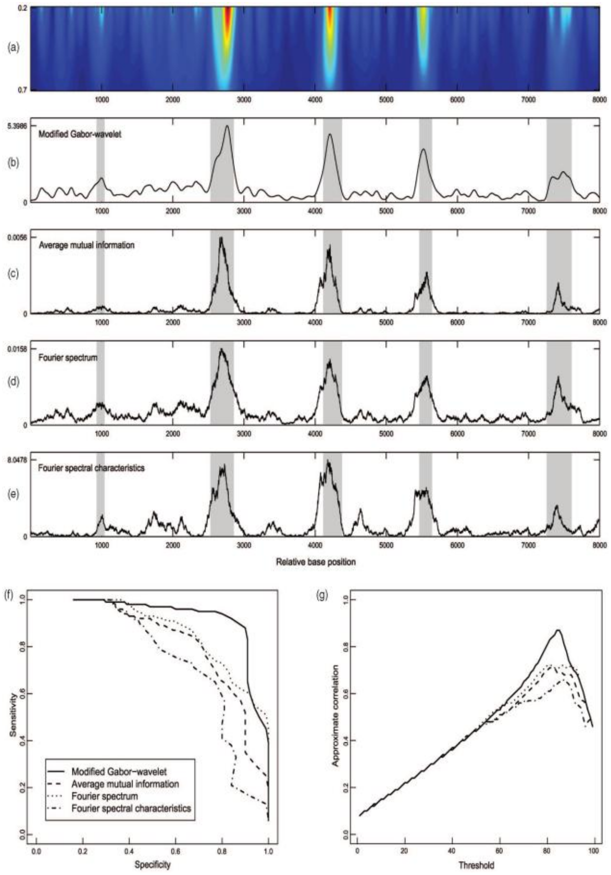


Figure 9. Identification of coding regions in F56F11.4 sequence using modified Gabor wavelet

The Discrete Wavelet Transform (DWT) is a mathematical tool that can be used very effectively for non-stationary signal analysis [26]. The DWT, for which an algorithm called Fast Wavelet Transforms (FWT) allows a very efficient calculation. Methods based on a modified Gabor-wavelet transform (MGWT) for the identification of protein coding regions also exists. Analysis is carried out for F56F11.4 sequence of C. elegans and the resulting spectrogram is shown in Fig. 9a. This figure represents the total spectrum, i.e., the sum of all spectrum values of the binary sequences. Fig. 9b represents the projections onto the position axis of the spectrum values. Figs. 9c, 9d, and 9e show the measures obtained with the average mutual information, Fourier spectrum, and Fourier spectral characteristics methods, respectively. A window length of 351 bp and a rectangular window step of 1 bp were adopted [27] and [28].

E. Entropy

Entropy measures are another example of a signal processing concept that has been used in genomic sequence analysis. The concept of entropy is used in signal analysis as a measure of randomness. The first definition of the entropy of a discrete information source (producing a discrete sequence) was introduced by Shannon [29]. The Minimum Entropy Mapping (MEM) Spectrum is a spectrum for symbolic sequences. It is independent from the mapping between symbols and numeric values, it possesses several key properties for the spectral representation of symbolic sequences, and it can be applied with limited complexity to DNA sequences. In addition, it can be easily extended to a local frequency spectrum that can be used to characterize non stationary sequences. The MEM spectrum can be used to improve the performances of gene finding programs, to investigate short range (less than 1000 bp) hidden correlations, and to analyze long range correlations. More generally, the MEM spectrum is a symbolic spectrum which can be used to search for DNA repetitive structures, such as tandem and dispersed repeats, or which can be extended to analyze the correlation between larger entities, such as codons and amino acids [30]-[32].

V. Conclusion

The application of Digital Signal Processing in DNA Sequence Analysis has received great attention in the last few years. This provides solution of various problems of living beings. Main DSP tools that have been applied in analysis are addressed. A recent development closely related to the impact of DSP on Bioinformatics is the new field of Genomic Signal Processing (GSP). This use DSP methods to obtain information from genomic and proteomic data to build models of molecular biological systems. Hence deeper understanding of the structure and functions of living systems will be obtained. This help in developing new diagnostic tools, therapeutic procedures and pharmacological drugs for applications like cancer classification and prediction [33].

References

[1] D. Anastassiou, "Genomic signal processing," *IEEE Sign Proc Mag*, vol. 18, no. 4, pp. 8-20, 2001.

[2] D. Anastassiou, "DSP in genomics processing and frequency-domain analysis of character strings," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing,* 2001, pp. 1053-1056.

[3] J. V. Lorenzo-Ginori, A. Rodriguez-Fuentes, R. G. Abalo, and R. S. Rodriguez, "Digital signal processing in the analysis of genomic sequences," *Current Bioinformatics*, vol. 4, pp. 28 – 40, 2009.

[4] D. L. Brutlag, "Understanding the human genome," in *Scientific American: Introduction to Molecular Medicine*, P. Leder, D. A. Clayton, and E. Rubenstein, Eds., New York NY: Scientific American Inc. 1994, pp. 153-168,.

[5] A. Khare, A. Nigam, and M. Saxena, "Identification of DNA sequences by signal processing tools in protein-coding regions," *Search & Research,* vol. 2, no. 2, pp. 44-49, 2011.

[6] J. Tuqan and A. Rushdi, "A DSP approach for finding the codon bias in DNA sequences," *IEEE J Select Topics Sign Proc,* vol. 2, pp. 343- 356, 2008.

[7] R. K. Deergha and M. N. S. Swamy, "Analysis of genomics and proteomics using DSP techniques," *IEEE Transactions on Circuits and systems*—I: Regular papers, vol. 55, no. 1, pp. 370-379, 2008.

[8] E. N. Trifonov, "3-, 10.5-, 200- and 400-base periodicities in genome sequences," *Physica A* , vol. 249, pp. 511-516, 1998.

[9] D. Kotlar and Y. Lavner, "Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions," *Genome Res* , vol. 13, pp. 1930-1937, 2003.

[10] T. W. Fox and A. Carreira, "A digital signal processing method for gene prediction with improved noise suppression," *EURASIP J Appl Sign Proc*, vol. 1, pp. 108-111, 2004.

[11] S. Datta and A. Asif, "DFT based DNA splicing algorithms for prediction of protein coding regions," in *Proc. IEEE Conference Record of 38th Asilomar Conference on Signals, Systems and Computer*, 2004, vol. 1, pp. 45-49.

[12] P. P. Vaidyanathan and B. J. Yoon, "The role of signal-processing concepts in genomics and proteomics," *J Franklin Inst* , vol. 341, pp. 111-35, 2004.

[13] J. Tuqan and A. Rushdi, "A DSP perspective to the period-3 detection problem," in *Proc. IEEE International Workshop on Genomic Signal Processing and Statistics*, 2006, pp. 53-54.

[14] A. Rushdi and J. Tuqan, "Trigonometric transforms for finding repeats in DNA sequences," in *Proc. IEEE International Workshop on Genomic Signal Processing and Statistics*, 2008, pp. 1-4.

[15] S. S. Sahu and G. Panda, "A DSP approach for protein coding region identification in DNA sequence," *International Journal of Signal and Image Processing*, vol. 1, no. 2, pp. 75-79, 2010.

[16] J. G. Proakis and D. K. Manolakis, *Digital signal Processing*, 4th Edition, Prentice Hall, NY 2006.

[17] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," *CABIOS* ; vol. 13, pp. 263-270, 1997.

[18] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*, Prentice-Hall, NY 2005.

[19] R. Romulus and G. Cornellia. Signal Processing Methods used in the Field of Genomics. [Online]. Available: http://www.upm.ro/InterIng2007/Papers/Section4/9-Reiz-Gordan%20Lucrare_pIV-9-1_4.pdf

[20] J. Epps, E. Ambikairajah, and M. Akhtar, "An integer period DFT for biological sequence processing," in *Proc. IEEE International Workshop on Genomic Signal Processing and Statistics*, 2008, pp. 1-4.

[21] A. Rodr ǵuez-Fuentes, J. V. Lorenzo-Ginori, and R. Grau-Ábalo, "Detection of coding regions in large DNA sequences using the short time fourier transform," *Lect Notes Comput Sci*, vol. 4225, pp. 902-909, 2006.

[22] P. P. Vaidyanathan and B. J. Yoon, "Gene and exon prediction using all pass based filters," *Workshop on Genomic Signal Processing and Stat., Raleigh, NC*, 2002.

[23] N. Chakravarthy, A. Spanias, L. D. Iasemidis, and K. Tsakalis, "Autoregressive modeling and feature analysis of DNA sequences," *EURASIP J Appl Sign Proc.*, vol. 1, pp. 13-28, 2004.

[24] M. Akhtar, E. Ambikairajah, and J. Epps, "Comprehensive autoregressive modeling for classification of genomic sequences," in *Proc. IEEE 6th International Conference on Information, Communications & Signal Processing*, 2007, pp. 1-5.

[25] H. X. Zhou, L. P. Du, and H. Yan, "Detection of tandem repeats in DNA sequences based on parametric spectral estimation," *IEEE Transactions on Information Technologyin Biomedicine*, vol. 13, no. 5, pp. 747-756, 2009.

[26] C. S. Burrus, R. A. Gopinath, and H. Guo, *Introduction to Wavelets and Wavelet Transforms: A Primer*, Prentice-Hall, NY 1997.

[27] P. Li ò, "Wavelets in bioinformatics and computational biology: state of art and perspectives," *Bioinform Rev*, vol. 19, pp. 2-9, 2003.

[28] J. P. Mena-Chalco, H. Carrer, Y. Zana, and R. M. Cesar Jr., "Identification of protein coding regions using the modified gabor-wavelet transform," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, no. 2, pp. 198-208, 2008.

[29] C. E. Shannon, "A mathematical theory of communication," *The Bell Sys Techn J*; vol. 27, pp. 379-423, pp. 623-656, 1948.

[30] W. A. Thompson, A. Martwick, and J. K. Weltman, "Decimative multiplication of entropy arrays, with application to influenza," *Entropy*, vol. 11, no. 3, pp. 351-359, 2009.

[31] L. G. R. Garello, "The minimum entropy mapping spectrum of a DNA sequence," *IEEE Transactions on Information Theory*, vol. 56, no. 2, pp. 771-784, February 2010.

[32] W. A. Thompson, A. Martwick, and J. K. Weltman, " Examining H1N1 through its information entropy," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 171-174, 2010.

[33] V. Varadan, P. Mittal, C. J. Vaske, and S. C Benz, "The Integration of biological pathway knowledge in cancer genomics, a review of existing computational approaches," *IEEE Signal Processing Magazine*, vol. 29, no. 1, pp. 35 – 50, 2012.

Inbamalar T M obtained her Bachelor's degree in Electronics and Communication Engineering from Bharathidasan University, India. Then she obtained her Master's degree from Madurai Kamaraj University. Currently, she is pursuing PhD in Information and Communication Engineering majoring in Genomic Signal Processing in Anna University of Chennai, India. She is currently working as Associate Professor in R M K Engineering College affiliated to Anna University, Chennai. Her specializations include Genomic signal processing, Digital Signal Processing and Digital Image processing. Her current research interests are Genomic and proteomic sequence analysis using Digital Signal Processing and Digital Image processing techniques.

Dr. **Sivakumar R** received his B.E degree in Electronics and Communication Engineering from Bharathiyar University, M.E degree in Medical Electronics from Anna University, Chennai, and the Ph.D. degree from the Anna University, Chennai. He is currently the Professor and Head in the Department of Electronics and Communication Engineering at R M K Engineering College. His research interests are Medical Signal and Image Processing and VLSI design. He has been invited to chair and speak at various International conferences all over the world. He is a life member of the Indian society of technical education. He is a member of the IEEE and senior member of IACSIT.