ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

DSP Applications for Genomic Sequences

Mangesh D. Ramteke[1], Nilima U. Rangari[2], Arun Katara[3]

Assistant Professor, Department of Electronics Engineering, Datta Meghe Institute of Engineering, Technology & Research, Sawangi(M), Wardha, India India[1]

Assistant Professor, Department of Botony, Nevjabai Hitkarini College, Bramhapuri, India[2]

Assistant Professor, Department of Electronics Engineering, Datta Meghe Institute of Engineering, Technology & Research, Sawangi(M), Wardha, India[3]

ABSTRACT: Digital Signal Processing (DSP) techniques are being used for the analysis as biological sequences form discrete nature. This paper attempt to develop effective DSP based techniques for biological sequence analysis. The main problem with gene detection approach is that it fails for certain genes. Analysis of genomic sequences requires defining an adequate representation of the nucleotide bases by numerical values. Thereafter, mathematical tools employed in DSP are used in identification of protein coding regions in DNA sequence, detection of the recurring patterns in these sequences, and others. Signal processing methods have played an important role in this context, some of which are reviewed in this paper. We describe some of the recent applications of mentioned DSP technique in the study of proteins having some comparison.

KEYWORDS- Digital Signal Processing, DNA sequences, Discrete Fourier Transform (DFT), digital filtering, Discrete wavelet transform (DWT).

I. INTRODUCTION

The detection of segments of DNA using computationally efficient algorithms is become a major area of research in computational biology. The primary approach in most cases is to look for specific characteristics or particular structure in the protein coding regions. DSP provides the information about the digital signals which are represented by a sequence of numbers. However, the genomic sequences can be represented mathematically by character strings of symbols from 4 alphabet sequences consisting of the letters A, T, G and C. As genomic represented by character string rather than numerical sequence, DSP doesn't show significant impact. However, if character strings are properly mapped into numerical sequences, DSP acts as a powerful tool for solving the related problems. The possibility of finding a wide application of DSP techniques to the analysis of genomic sequences arises when these are converted appropriately into numerical sequences, for which several rules have been developed. Notice that genomic signals do not have time or space as the independent variable, as occur with most physical signals.

The Digital Signal Processing (DSP) techniques uses a set of mathematical tools to analyze and process signals, among them can be mentioned as Discrete Fourier Transform, Z transform, Digital Filters, Parametric Models, the Wavelet Transform, Correlation Functions and few more. For example, measurement of latent Periodicity of biological sequence to identify gene locations or structure within sequences. The discrete Fourier transform (DFT) has received considerable attention as a tool for detecting such periodicities over shorter sequence segments. Correlation Functions are used to characterize such latent periodicities.

This paper is organized in a manner wherein overview of basic concepts of molecular biology followed by the various techniques used for numerical representation of genomic sequences is described. This helps us to study various applications of DSP tool in regards with genomic sequence. We will also review various DSP algorithms used in genomic sequence analysis such as digital filters, the Discrete Fourier Transform (DFT), Discrete Wavelet Transform and the Information Theory concept of entropy.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

II. RELATED WORK

DSP methods that described period-three behaviour have many shortcomings. These methods are unable to locate coding regions that do not have strong period-three characteristics, for which hidden Markov models [1] provides good results in these regards. Specific structure of DNA consists of four nucleotides which are designated by the characters A, T, C, and G. A character string composed of these four nucleotides can be mapped to four signals [16]. In a similar fashion, the DFT can be obtained. For many genes, period-three behaviour has been observed and is useful for identifying coding regions [17]. Specifically, the (k = N/3)-DFT coefficient magnitude is often significantly larger than the surrounding DFT coefficient magnitudes and corresponds to a coding region within the gene [16,17]. High frequency selective bandpass digital filters for the identification of coding regions can be used instead of the DFT. The digital filter described in [2] is a second order anti notch filter. The digital filter method for the identification of coding regions does not require the use of a sliding window [2, 11] by Vaidyanathan and Yoon.

III. BIOLOGICAL CELLS: FUNDAMENTAL BUILDING BLOCKS

All living organisms are made up of microscopic fundamental biological structures called cells. Living cells are classified into two types, the simpler prokaryotic cell and eukaryotic cell. An overview of some of the important aspects of the DNA molecule, RNA molecule and proteins is described as follows.

A. *DNA*

DNA is made up of molecules called nucleotides. Each nucleotide contains a phosphate group, a sugar group and a nitrogen base. The four types of nitrogen bases are adenine (A), thymine (T), guanine (G) and cytosine (C). The order of these bases is what determines DNA's instructions. Similar to the way the order of letters in the alphabet can be used to form a word, the order of nitrogen bases in a DNA sequence forms genes, which in the language of the cell, tells cells how to make proteins.

Fig. 1 shows a helix structure of a DNA molecule. 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. DNA sequencing is technology that allows researchers to determine the order of bases in a DNA sequence. The technology can be used to determine the order of bases in genes, chromosomes, or an entire genome.

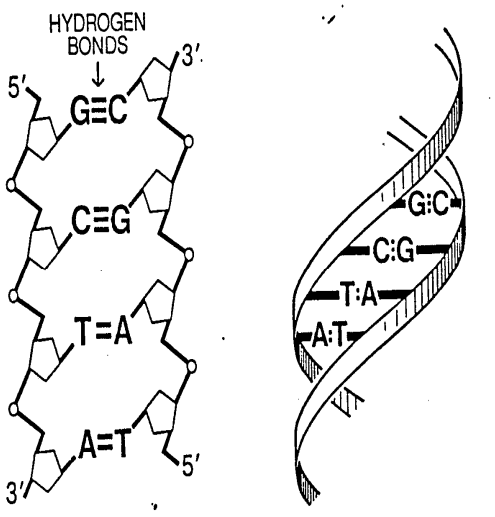

Fig 1: Helix Structure of DNA

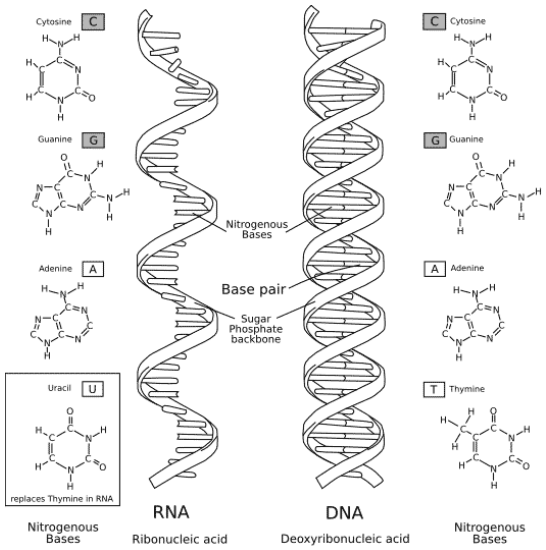

Fig 2: Structure of RNA

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

B. *RNA*

Ribonucleic acid (RNA) is a working copy of DNA resulting from a process known as transcription based on the information contained in DNA. RNA is very similar to DNA except that in RNA the nucleotide uracil(U) replaces thymine (T) in DNA, and RNA is normally found as a single-stranded molecule, whereas DNA is double stranded. From the viewpoint of genetic information, T in DNA and U in RNA are equivalent. The main job of RNA is to transfer the genetic information contained in DNA from nucleus to ribosome for the creation of proteins. This process prevents the DNA from having to leave the nucleus. This process keeps the DNA and genetic code protected from being corrupted. Fig. 2 shows chromosome containing DNA molecule and the process of RNA.

C. *Proteins*

Proteins are made up of hundreds or thousands of smaller units called amino acids, which are attached to one another in long chains. There are 20 different types of amino acids that can be combined to make a protein. The sequence of amino acids determines each protein's unique 3-dimensional structure and its specific function. 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 consists of 64 triplets of nucleotides. These triplets are called codons. With three exceptions, each codon encodes for one of the 20 amino acids used in the synthesis of proteins. That produces some redundancy in the code, most of the amino acids being encoded by more than one codon. Knowledge of the genetic code allows one to predict the amino acid sequence of any sequenced gene. The complete genome sequences of several organisms have revealed genes coding for many previously unknown proteins.

Table I: The Genetic Code

1st	2nd								3rd
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
	UUA	Leu	UCA	Ser	UAA	Term	UGA	Term	A
	UUG	Leu	UCG	Ser	UAG	Term	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	AUG*	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GUG*	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

Position in Codon

IV. SEQUENCE REPRESENTATION OF DNA

DNA has double stranded anti-parallel helix. It is built by concatenating nucleotides namely Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). Adenine (A) on one strand always binds with a Thymine (T) on the other strand, Also Cytosine (C) always binds with a Guanine (G). Therefore, it posses Complementary property between DNA double-strands. The bases of a DNA sequence are to be mapped onto their corresponding numerical values in order to apply digital signal processing. The nucleotides of DNA data can be represented into a series of arbitrarily numerical sequences such as 2-bit binary, the 4-bit binary, the paired nucleotide, and the 12-letter alphabet representations. Each of the DNA numerical representations attains various properties, and maps a DNA sequence into one to twelve numerical sequences.

The nucleotides A, C, G, and T can mapped into two-bit binary i.e. 00, 11, 10, 01 respectively which shows 1-dimensional indicator sequence. For 4-bit binary encoding, the nucleotides A, C, G, and T are mapped as 1000, 0010, 0001 and 0100 respectively which also show a 1-dimensional indicator sequence. Genetic code context (GCC)

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

representation involves the composition and distribution of the amino acid information in three reading frames. In this method, each consecutive codon from the three reading frames in the DNA sequence is converted to an amino acid and each amino acid in turn is represented by a unique complex number. This results in a single dimension indicator sequence in amino acid domain.

V. DSP APPLICATIONS FOR DNA ANALYSIS

The analysis of DNA sequences using DSP can be useful in the detection of protein coding regions in genomic sequences. 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.

A. *Discrete Fourier Transform*

The DFT is the most important discrete transform, used to perform Fourier analysis in many practical applications. In digital signal which enables us to find the spectrum of a finite-duration signal with the help of following definition.

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k)\, e^{\frac{j2\pi kn}{N}} \qquad n = 0,1,\ldots..,N-1$$

Discrete Fourier Transform evaluates the frequency components required to reconstruct the finite segment of the sequence that was analyzed. The Discrete Fourier transform can reveal periodicities in the input data as well as the relative intensities of these periodic components. Fig. 3 shows a spectrum obtained by taking DFT for a DNA sequence.

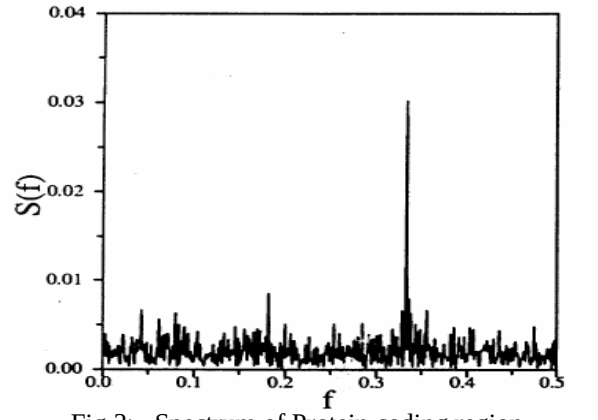

Fig 3: Spectrum of Protein coding region

When the DFT is used for signal spectral analysis, the x{n}, sequence usually represents a finite set of uniformly spaced time-samples of some signal x(t), where t represents time. To achieve a statistically good result, the DFT for spectral analysis of random signals require certain considerations. For stationary random signals, Welch's modified periodograms function method is commonly used to obtain a power spectral density (PSD). 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. For non-stationary signals, The Short Time Fourier Transform (STFT) is frequently used for the DFT-based spectral analysis as shown in Fig. 4. Here, the signal is divided into short segments and a DFT is calculated for each one of these segments. A three dimensional plot i.e. spectrogram of the squared magnitude of the DFT coefficients against time is obtained.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

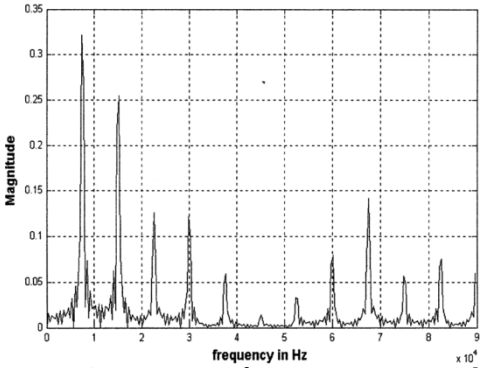

Fig 4: Spectrogram of non-stationary signal

B. *The Discrete Wavelet Transform (DWT)*

In recent years, wavelets analysis has been applied to a large variety of biomedical signals, wherein it is found a growing interest in using wavelets in the analysis of sequence and functional genomics data. Therefore, this review is intended to give a relatively accessible introduction to wavelet analysis. In wavelet analysis, the Discrete Wavelet Transform (DWT) decomposes a signal into a set of mutually orthogonal wavelet basis functions. The wavelet transform basis functions are compact, or finite in time, while the Fourier sine and cosine functions are not. This feature allows the wavelet transform to obtain time information about a signal in addition to frequency information. Fig. 5 depicts this relationship by showing how the time resolution gets finer as the scale/frequency increases.

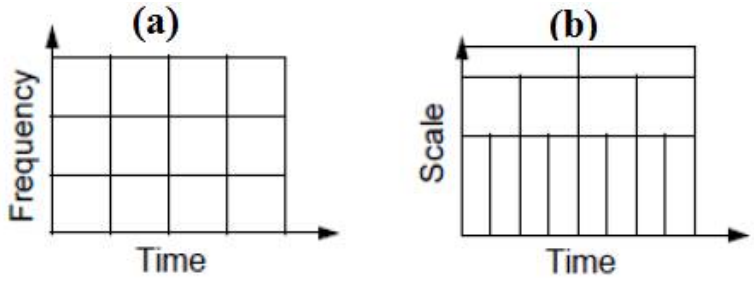

Fig 5: (a) STFT and (b) DWT Frequency/Time

Wavelets can be useful in detecting patterns in DNA sequences. Wavelet variance decomposition can also be applied for bacterial genome as described by Lio. Gabor-wavelet transform (MGWT) for the identification of protein coding regions is also proposed for specific patterns of nucleotides at coding regions.

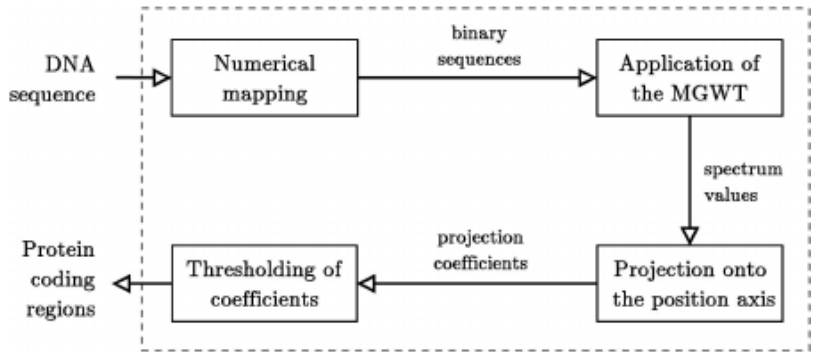

Fig 6: Schematic data flow diagram using the MGWT

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

C. *Parametric Spectral Analysis*

Spectral analysis of DNA gene expressions time series data is important for understanding the regulation of gene expression and gene function. The cell cycle regulation of gene expression profiles based on the combination of singular spectrum analysis (SSA) and autoregressive (AR) spectral estimation. Using the SSA, the dominant trend of data can be extracted and reduce the effect of noise. Based on the AR analysis, high resolution spectra can be produced.

D. *Digital Filters*

An efficient way to improve the detection of periodicities in a DNA sequence is digital filtering. The filtering can be accomplished by real FIR low-pass filters that are easy to design wherein DNA spectrum can be effectively calculated without computing the DFT. There are two basic types of digital filters, Finite Impulse Response (FIR) and Infinite Impulse Response (IIR) filters. The general form of the digital filter difference equation

$$y(n) = -\sum_{k=1}^{N} a_k \, y(n-k) + \sum_{k=0}^{M} b_k \, x(n-k)$$

Here, a_k and b_k are the coefficients of the filter. Another way to characterize a Filter is as a system function which is a ratio of two polynomials in z^{-1}.

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{k=0}^{M} b_k z^{-k}}{1 + \sum_{k=1}^{N} a_k z^{-k}}$$

An associated frequency response can be obtained by mapping as in following equation.

$$H(e^{j\omega}) = H(z)\big|_{z=e^{j\omega}}$$

A variety of digital filter design techniques allow to obtain desired magnitude response with frequency selectivity properties, whereas it is desired that the phase response be a linear function of ω, in order to have low distortion. These ideal responses can be approximated to the desired responses, where better approximations in general are obtained by increasing the order of H(z).

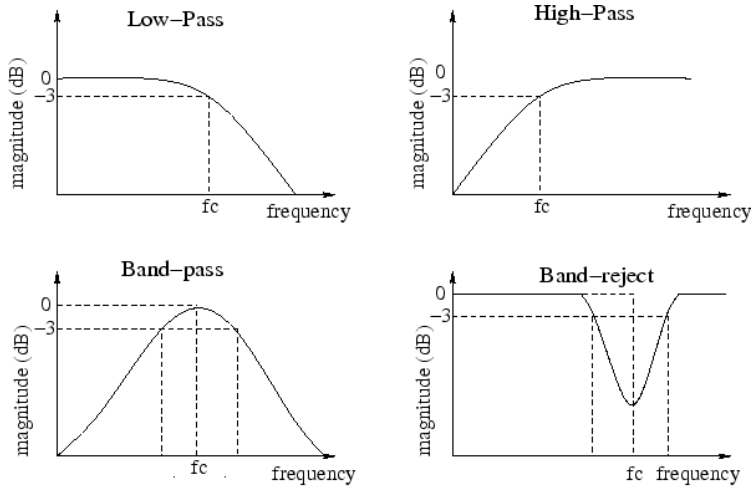

Fig 7: Frequency response in magnitude of ideal filters

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 3, March 2018

VI. CONCLUSION

The DSP applications in Genomic Sequence Analysis have received extensive attention in the last few years which suggests much solution of various problems of organisms, botanical study as well. A recent application development makes an impact of Digital Signal Processing in the field of Genomic Analysis. All mentioned DSP approach to obtain information from genomic sequences is used to build models of various molecular biological systems. Thereby provides an understanding of the structure and various functions of living organisms. This may help in medical diagnostic and remedies for various malfunctions.

REFERENCES

[1] J. Henderson, S. Salzberg, and K. H. Fasman, "Finding genes in DNA with a hidden Markov model," J. Comput Biol., vol. 4, no. 2, pp. 127–141, 1997.
[2] P. P. Vaidyanathan and B.-J. Yoon, "Gene and exon prediction using allpass-based filters," in *Workshop on Genomic Signal Processing and Statistics*, Raleigh, NC, USA, October 2002.
[3] 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.
[4] Swarna bai Arniker, Hon Keung Kwan, "Advanced Numerical Representation of DNA Sequences," in 2012 International Conference on Bioscience, Biochemistry and Bioinformatics, IPCBEE vol.31(2012) © (2012)IACSIT Press, Singapoore.
[5] P. Lio, and M. Vannucci. Finding pathogenicity islands and gene transfer events in genome data. Bioinformatics, 2000, 16(10):932-940.
[6] J. A. Berger, S. K. Mitra, and J. Astola, "Power spectrum analysis for DNA sequences", Proc. of Seventh International Symposium on Signal Processing and its Applications, 2003, 2:29-32..
[7] R. Ranawana and V. Palade. A Neural network based multi-classifier system for gene identification in DNA sequence. Neural Computing and Applications,2005,14(2):122-131.
[8] J. B. Demeler, G. W. Zhou. Neural network optimization for E.coli promoter prediction. Nucleic Acids Res., 1991, 19(7):1539-1599.
[9] C. Yin, S. Yau. Numerical representation of DNA sequences based on genetic code context and its applications in periodicity analysis of genomes. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 2008
[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] 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.
[14] https://www.livescience.com/37247-dna.html.
[15] J. G. Proakis and D. K. Manolakis, Digital signal Processing, 4th Edition, Prentice Hall, NY 2006.
[16] 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.
[17] D. Anastassiou, ―Genomic signal processing,‖ IEEE Sign Proc Mag, vol. 18, no. 4, pp. 8-20, 2001.
[18] 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.
[19] 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.
[20] 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.
[21] P. Lio, "Wavelets in bioinformatics and computational biology: state of art and perspectives", Bioinform Rev, vol. 19, pp. 2-9, 2003.
[22] 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.
[23] 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, 2008.
[24] 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.