International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 04 | Apr -2017 www.irjet.net p-ISSN: 2395-0072

Analysis of Genomic sequences for prediction of Cancerous cells using Wavelet technique

T.Thillai Gayathri

PG Scholar, Dept. Of ECE,PSG College of Technology, Coimbatore, Tamilnadu, India.

Abstract - *Nowadays, the researchers have been faced many challenges in analyzing the vast DNA sequences. In analysis of enormous amount of data, Signal Processing concepts especially Wavelet Transform techniques play an important role. One of the main reasons for deadly disease like cancer is the Genetic abnormality. The abnormal changes in the genes (mutation of DNA sequences) can cause cancer. The main aim of this paper is to identify the major changes occur in the normal sequences due to mutation. In this paper, the EIIP representation technique is used which has major advantage of reducing computational overhead compared to other representation techniques. Here, MATLAB R2014a is used which supports bioinformatics toolbox. The DNA sequences have been collected from NCBI website for analysis.*

Key Words: Deoxy Ribo Nucleic Acid (DNA), Digital Signal Processing *(DSP), Discrete Wavelet Transform (DWT), Genomic Signal Processing(GSP), Electron Ion Interaction Potential(EIIP).*

1.INTRODUCTION

Genome Signal Processing [1] is an emerging technology in the research of deadly disease like cancer nowadays. Analysis of genes or genomic data by applying Digital Signal Processing techniques is known as GSP. Bioinformatics [2] is an interdisplinary field which is the combination of biology with computer science. This technology describes the phenomenon of the disease at the molecular level (gene level). It is the technology which improves the accuracy of the result, reduces the time of drug discovery and cost effective.

Cancer, a malignant neoplasm (in medical term) is a deadly disease which is caused due to mutation (alteration) in DNA sequences [3]. Nowadays, mortality rate has been increased due to cancer. During Cell division, the cell gets divided to make new ones as human body grows. Cancer is also caused due to some environmental agents such as exposure to chemicals or radiations. According to scientific discoveries [4], DNA plays a major role in the study of cancer. Genomic data such as DNA basically in character form (A, T, G, C) converted into digital form which is known as Genomic signal. In analyzing the DNA sequences, it is mandatory to convert the biological sequences into numerical sequences. There are so many mapping techniques [5] are available for performing conversion of DNA sequences. Some of them are Integer representation, Voss representation, Tetrahedron representation, Complex representation, etc…

In [6], Abo-Zahhad explains the various types of conversion techniques clearly with its merits and demerits. Voss representation [7] is one among the numerical way of representation which converts one DNA sequence into four binary sequences. Since the result of this representation becomes four sequences, it increases computational complexity. To overcome this drawback, EIIP representation technique have been introduced. It reduces computational overhead by 75% [8] by applying corresponding EIIP values to each nucleotide (A, T, G, C). It exhibits periodicity property and also improves differentiation or discrimination capability of genes.

The important factor in analysis of genomic sequences is to identify the protein coding region [9]. In [10], Anastassiou demonstrates the identification of protein coding and non coding regions by employing DSP techniques to the numerical values of the DNA strings. In [11], Jianchang Ning explains the wavelet applications in analyzing biological problems .The main point that they conclude in this paper is the wavelet technique, which is the best way of analyzing biological sequences. Wavelets provides the signals in multiscale way of representation.

The paper is organized as follows. Section I describes the introduction part. Sections II presents the Molecular biology concepts and techniques of signal processing. Section III introduces the proposed methodology. Section IV depicts the algorithm of the proposed method. In Section V, the results and simulation by using algorithm of the proposed method is reviewed, which clearly shows the discrimination between the normal and the mutated genes. And Finally, section VI concludes the paper.

2. MOLECULAR BIOLOGY AND SIGNAL PROCESSING CONCEPTS

2.1 Genome and DNA

Human body are made up of numerous of cells. In order to maintain life, all these cells are working together properly. The DNA is found in the nucleus of every cell in the human body. It looks like a twisting ladder. The DNA [12] is made up of two sequences of nucleotides and hence it is said to be double stranded molecule. The double stranded DNA molecule consists of four nucleotides such as Adenine (A), Guanine (G), Cytosine (C), Thymine (T). These nucleotides are tightly bounded or connected together by chemical bonds. A always pairs with T, G is always bounded to C. For example, the chemical bond always exists between the nucleotide Adenine(A) in the strand 1and Thymine(T) in the strand 2 and also Guanine(G) in the strand 1 and Cytosine (C) in the strand 2.

Strand 1: G – A – C – G – A – T – A – G

Strand 2: C – T – G– C – T – A – T– C

The structure of DNA and its location are shown in fig.1.

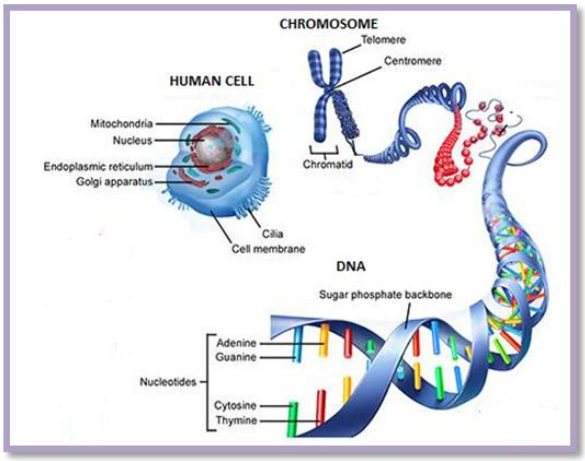

Figure -1: DNA Structure and its Location

The mutation in Genes [13] can cause cancer disease. During Cell division, the DNA may gets mutated. The cells divides to make new cells as human body grows. The special areas of DNA molecule regulates during this cell division. The basis of tumour can be formed by mutation of DNA molecule in some of the regulatory areas. The effects of mutation is shown in the fig.2.

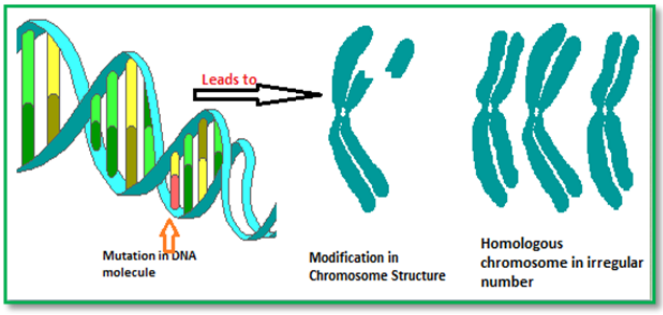

Figure -2: Effects of Mutation

Various types of mutation can be occurred in the DNA molecule [14,15] are shown in the fig.3. The types of mutations are, 1.Substitution, 2.Insertion or Deletion (Indel), 3.Copy number alterations, and 4.Translocations.The change of one nucleotide in the DNA sequence leads to alteration in the protein sequence due to change of amino acid which leads to the functional changes in the protein also. One nucleotide is substituted by another nucleotide is called substitution. The insertion or deletion of nucleotide in the DNA sequence is called Indel mutation. In the third type of mutation (copy number of alteration), the amplification (increase) of genes and deletion (decrease) of genes are possible. Fourth mutation called translocation leads to change in the physiology of normal cells, over expression of a gene and also cause pathogenesis of cancer.

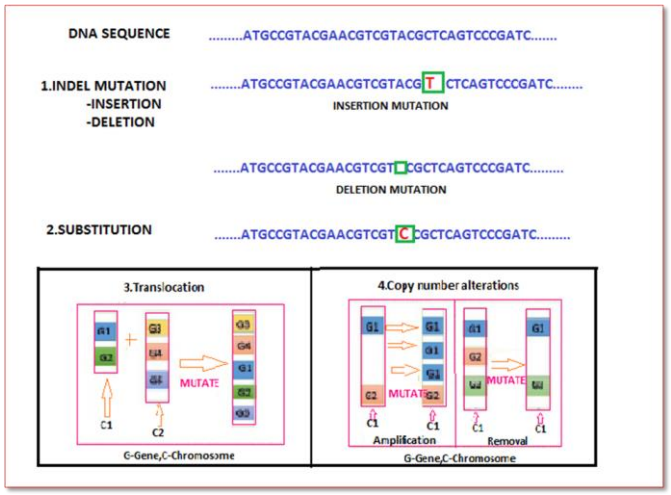

Figure -3: Mutation Types

2.2 Signal Processing Concepts in analysis of DNA

Digital Signal Processing (DSP)[16] concepts and its applications in Bioinformatics have attained great attention in recent years, where new effective methods for analysis of genomic sequences such as the detection of coding regions, have been developed. Defining an adequate representation of the nucleotide bases by numerical values is the major requirement for genomic sequence analysis by the use of DSP concepts and its principles. After performing conversion technique, the DSP concepts have been applied for further analysis. The application of signal processing techniques have been achieved more advantages than the application of biological experiments in analyzing biological sequences.

The traditional Fourier Transform (FT) is one of the DSP concepts. But it has the limitation of analyzing both time and frequency domain simultaneously. And also it is not suitable for analyzing non-stationary signals. The Wavelet transform is the most predominant technique, which is used for analyzing both stationary and non-stationary signals. Another main advantage of wavelet transform is its independent of window length [17,18].

3. METHODOLOGY

According to the project layout, the steps involved in the proposed methodology are Collection of Databases, Conversion of nucleotide sequences into numerical sequences, Analysis of Wavelet techniques and so on. The layout of the proposed method is shown in the fig.4

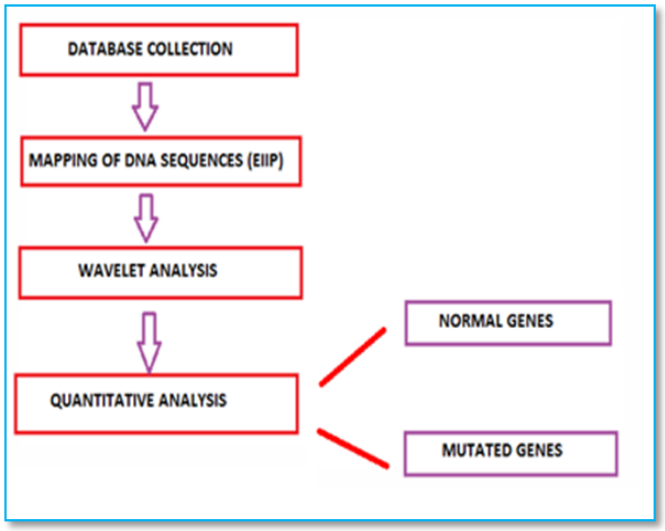

Figure -4: Layout of the Proposed Method.

3.1 Collection of Databases

The first and foremost step is to collect the required databases of DNA sequences. Here, it is to be collected from the NCBI [19] website. It is an open access website so that it is very easy for collecting the DNA sequences for further analysis. Each sequence contains more than 1000 nucleotides in the form of A,T,G,C.(Adenine, Thymine, Guanine, Cytosine).

3.2 Number and Mapping

After collecting the required databases, it is mandatory to convert those symbolic sequences into numbers. For conversion process, EIIP representation technique [20] is used here for several benefits. It reduces computational complexity. It provides biological information without much loss of information. For example, if the DNA sequence x[n]=[…..GTCATGCTGACGT…….],then the numerical conversion sequence using EIIP representation technique is [……0.0806 0.1335 0.1340 0.1260 0.1335 0.0806 0.1340 0.1335 0.0806 0.1260 0.1340 0.0806 0.1335……..] as mentioned in table[1].

Table -1: EIIP Value for nucleotides

Nucleotide	EIIP Value
A	0.1260
T	0.1335
C	0.1340
G	0.0806

For clear clarification, the EIIP plot for the sample DNA sequence is shown in the fig.5.

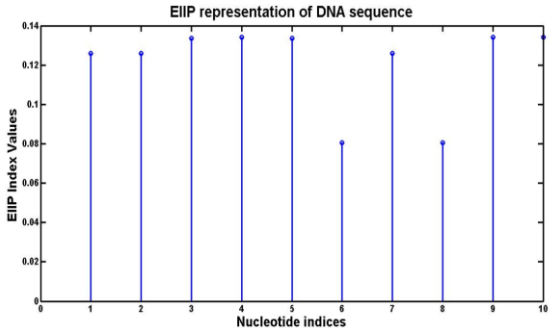

Figure -4: Sample EIIP plot for the DNA sequence AATCTGAGCCAAGTAGAAGAC….The x-axis represents nucleotide indices(A,T,C,G) and the y-axis represents EIIP values(0.1260,0.1335,0.1340,0.0806).

3.3 Analysis of Wavelet technique

Wavelet transform [21,22] is more better than other transform techniques in the application of biological signals since it is used for analyzing non stationary signals. This technique is independent of window length and it uses small scales for analyzing small coding regions and large scales for large coding regions. Here, Wavelet transform is applied for numeric DNA sequences to identify the mutation spots by performing various types of mutation in the normal DNA sequences. Wavelet transform has better resolution in both time and frequency due to its independent size of the window.

3.4 Evaluation

The performance metrics computed here are Mean amplitude, Standard Deviation, ratio between the mean amplitude and the standard deviation. The ratio which is less than 1 obtained for mutated sequences and more than 1 for normal sequences.

4. ALGORITHM

The following steps are the algorithm of proposed method.

1. The DNA sequences of normal cells are extracted from open access NCBI website.
2. The DNA sequences of symbolic representation can be converted into numerical sequences by using most advantageous representation technique called EIIP.
3. After converting into numerical sequences, mutate some base pairs of DNA sequences manually by performing various types of mutations which is clearly explained in section II.
4. Wavelet transform is applied for both normal and mutated sequences. It is used for analysing the DNA sequences without using the biological experiments because of the advantage of easily visualizing the variation in the sequences due to mutation.
5. The parameters such as standard deviation and mean amplitude are also calculated for clearly observing the variations between the sequences.

5. RESULTS AND ANALYSIS

The proposed method involves the identification of mutated spots. Mutation is done manually using the types of mutation discussed in section II. Wavelet transform is applied for both normal and the manually mutated sequences and variations due to mutation can be visualized. In addition to that, the mutated cells are identified by calculating the parameters like mean amplitude, standard deviation. The ratio of mean amplitude to standard deviation is more than 1 indicates normal cell and less than 1 indicates mutated cells [23].

Table -2: EIIP Value for nucleotides

S.No	Accession Numbers	Gene Name	Relative Position	Length of Exon
1	AF186611	HBB	987:1172	186
2	AF186613.1	HBB	987:1172	186
3	AF186607.1	HBB	988:1173	186
4	AF348448	HBB	139:324	186
5	AF186608	HBB	989:1175	186
6	AF186616	HBB	992:1177	186
7	AF083883	HBB	1237:1425	189
8	AF186614	HBB	988:1173	186

5.1 Result Analysis for Normal and Mutated Sequences

Before Mutation

The wavelet plots for normal sequences are as below,

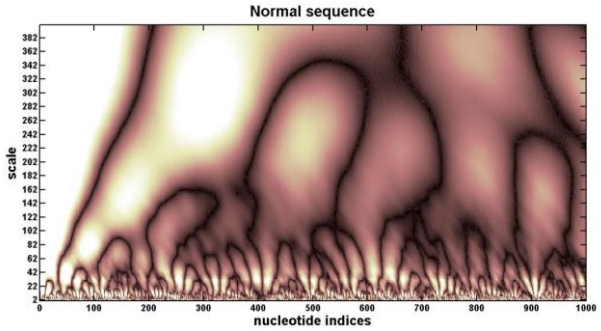

Figure- 6: Wavelet Transform Plot for Normal Cell of Accession no. AF186608

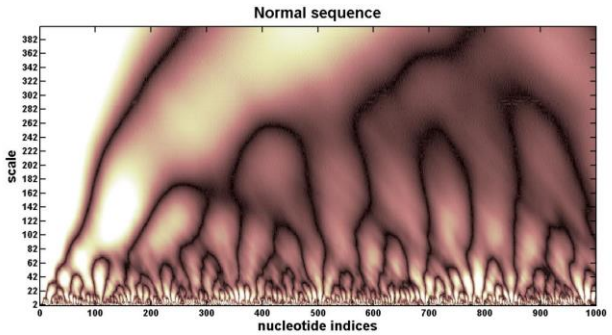

Figure- 7: Wavelet Transform Plot for Normal Cell of Accession no. AF186613.1

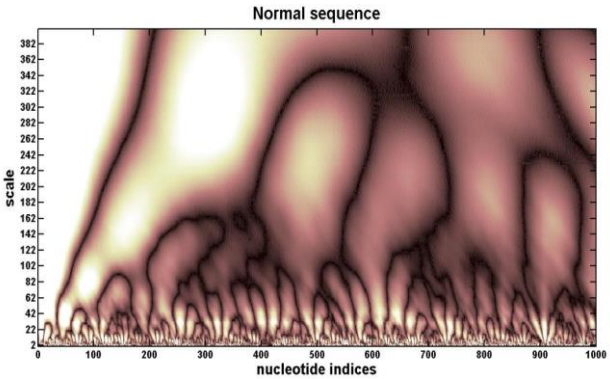

Figure- 8: Wavelet Transform Plot for Normal Cell of Accession no. AF083883

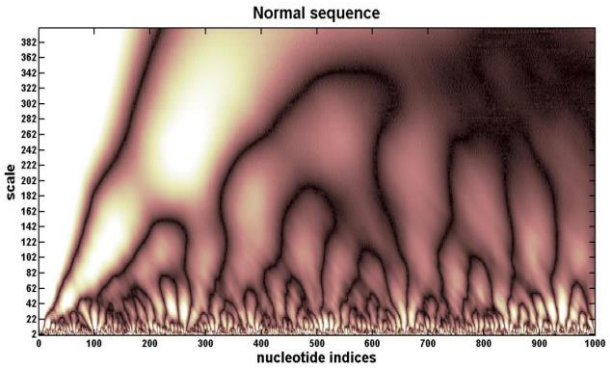

Figure- 9: Wavelet Transform Plot for Normal Cell of Accession no. AF186611

AFTER MUTATION

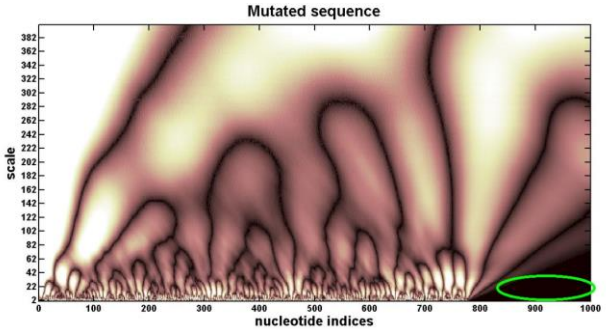

Figure-10: Wavelet Transform Plot for mutated cell of Accession no. AF186608

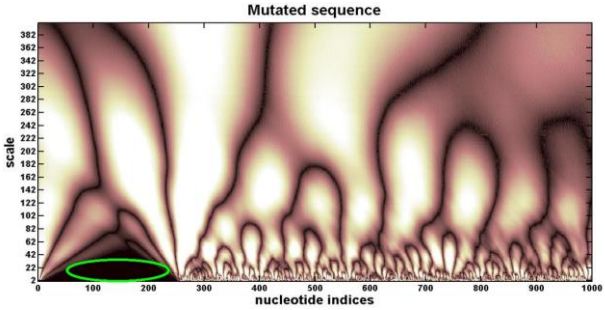

Figure-.11: Wavelet Transform Plot for Mutated Cell of Accession no. AF186613.1

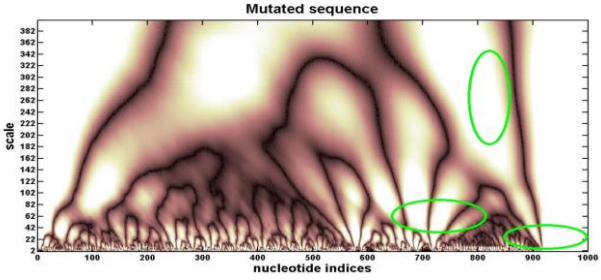

Figure-12: Wavelet Transform Plot for Mutated Cell of Accession no. AF083883

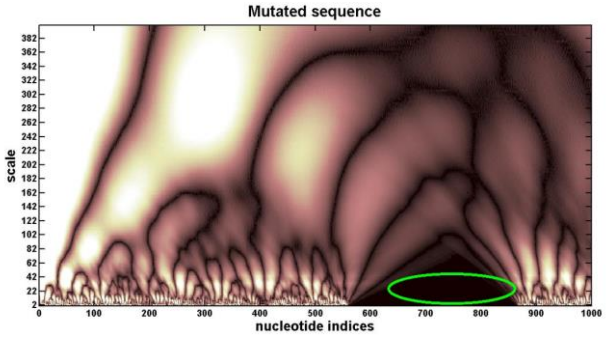

Figure-13: Wavelet Transform Plot for Mutated Cell of Accession no. AF186611

5.2 Result Analysis for Normal and Mutated Sequences

Table -3: Analysis of Normal Cells

S.No	Accession Numbers	Mean amplitude (X)	Standard deviation (Z)	Ratio (X/Z)
1	AF186608	0.0683	0.0321	2.1277
2	AF186613.1	0.0467	0.0292	1.5993
3	AF083883	0.0976	0.0806	1.2109
4	AF186611	0.1465	0.1291	1.1352

Table -4: Analysis of Mutated Cells

S.No	Accession Numbers	Mean amplitude (X)	Standard deviation (Z)	Ratio (X/Z)
1	AF186608	0.2367	0.3265	0.7249
2	AF186613.1	0.4327	0.8143	0.5313
3	AF083883	0.0567	0.0765	0.7411
4	AF186611	0.5478	0.8345	0.6564

Table [3] shows the parameters such as mean amplitude, Standard Deviation for both normal and mutated cells. The inference obtained from the above table is that ratio of mean amplitude to standard deviation is more than 1 for normal

cells and less than 1 for mutated cells. It is visualized via bar chart as shown in chart.1.

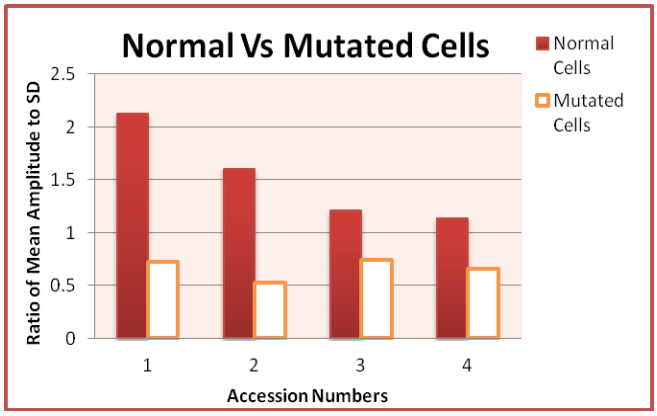

Chart -1: Ratio of Mean amplitude to standard deviation.

6.CONCLUSION

Nowadays, DSP plays a predominant role in analysing the DNA sequences without performing any biological experiments. In this paper, an efficient algorithm have been developed for analysing the DNA sequences in MATLAB (R2014a) environment which supports bioinformatics toolbox. The numerical representation technique used here is EIIP which has many more advantages compared to other representation techniques. Wavelet technique is applied for identifying the mutation spots. The biological experiments available for cancer prediction are costly whereas the Signal processing techniques consumes less time and cost effective. As the major cause for cancer disease depends on mutation in DNA sequences, wavelet based technique has great scope in future due to their attractive properties such as representation of time frequency domain, identification of local feature, multi-resolution capability, etc. Without performing biological experiments, one can easily predict the deadly disease like cancer by using signal processing concepts.

REFERENCES

[1] Dougherty Edward R. and Dutta Aniruddha: "Genomic Signal Processing: Diagnosis and Therapy " IEEE Signal Processing Magazine January, 2005.

[2] N.M. Luscombe, D. Greenbaum, M. Gerstein," Bioinformatics, An introduction and overview", Yearbook of Medical Informatics 2001.

[3] Michael R. Stratton, Peter J. Campbell and P. Andrew Futreal1, "The cancer genome," Nature, vol. 458, pp. 719-724, 9 April 2009.

[4] Tao Meng ,Shu-ching Chen, "Wavelet analysis in current cancer Genome Research: A Survey", IEEE /ACM transactions on computational biology and bioinformatics, vol. 10, pp. 567-570, 2013.

[5] P. D. Cristea, "Conversion of nucleotides sequences into genomic signals," Journal of Cellular and Molecular Medicine, vol. 6, pp.279-303, 4 April 2002.

[6] Mohammed Abo-Zahhad, Sabah M. Ahmed, Shimaa A. Abd-Elrahman,"Genomic Analysis and Classification of Exon and Intron Sequences Using DNA Numerical Mapping Techniques ", I.J. Information Technology and Computer Science,Vol. 8, pp. 22-36, 2012.

[7] R.F Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences", Physical Review Letters, Vol. 68, pp.3805- 3808, 1992.

[8] Mai S. Mabrouk, "A study of the Potential of EIIP mapping method in exon prediction using the frequency domain technique ,"American Journal of Biomedical Engineering, vol. 78, pp. 17-22, October 2012.

[9] Parameswaran Ramachandran, Wu-Sheng Lu, and Andreas Antoniou." Location of Exons in DNA Sequences Using Digital Filters",IEEE Transactions,pp.2337-2340,2009.

[10] D. Anastassiou, "Genomic Signal Processing." IEEE Signal Processing Magazine, vol.18, pp. 8-20, 4 April 2001.

[11] C. Gargour, M. Gabrea, V. Ramachandran, and J.M. Lina, "A Short Introduction to Wavelets and Their Applications," IEEE Circuits and Systems Magazine, vol. 9, no. 2, pp. 57-68, Second Quarter 2009.

[12] J.D. Watson & F.H.C. Crick, (1953) "A structure for DNA", Nature, Vol. 171, pp. 737-738, 25 April 1953.

[13] E.P. Reddy, R.K. Reynolds, E. Santos, and M. Barbacid, "A Point Mutation Is Responsible for the Acquisition of Transforming Properties by the T24 Human Bladder Carcinoma Oncogene," Nature, vol. 300, no. 5888, pp. 149-152, July 1981.

[14] T. Boveri, "Concerning the Origin of Malignant Tumours by Theodor Boveri," J. Cell Science, vol. 121, no. Supplement 1, pp. 1-84, 2008.

[15] X. Xu, K. Zhu, F. Liu, Y. Wang, J. Shen, J. Jin, Z. Wang, L. Chen, J. Li, and M. Xu, "Identification of Somatic Mutations in Human Prostate Cancer by RNA-Seq," Gene, vol. 519, no. 2, pp. 343-347, May 2013.

[16] P. P. Vaidyanathan and B. J. Yoon, "The role of signal processing concepts in genomics and proteomics," Journal of the Franklin Institute, Vol. 341, pp. 111-135

[17] M. Sifuzzaman, M.R. Islam, and M.Z. Ali, "Application of Wavelet Transform and Its Advantages Compared to Fourier Transform," J. Physical Sciences, vol. 13, pp. 121-134, 2009.

[18] A.K. Nagar and D. Sokhi, "On Wavelet-Based Adaptive Approach for Gene Comparison," Int'l J. Intelligent Systems Technologies and Applications, vol. 5, pp. 104-114, 2008.

[19] National Centre for Biotechnology Information (NCBI).Available: http://www.ncbi.nlm.nih.gov/.

[20] A. S. Nair and S. P. Sreenadhan, "A coding measure scheme employing electron-ion interaction pseudopotential (EIIP)," Bioinformation, vol. 1, no. 6, pp. 197–202, 2006.

[21] J.K. Meher, M.K. Raval, P.K. Meher, and G.N. Nash, "Wavelet Transform for Detection of Conserved Motifs in Protein Sequences with Ten Bit Physico-Chemical Properties," Int'l J. Information and Electronics Eng., vol. 2, no. 2, pp. 200-204, 2012.

[22] E. Lin and E. Linton, "Wavelet Packet Analysis of DNA Sequences," Proc. Fifth Int'l Conf. Bioinformatics and Biomedical Engg. (ICBBE), pp. 1-3, May 2011.

[23] Shilpi Chakraborty, Vinit Gupta, "Dwt based cancer identification using EIIP",International Conference on Computational Intelligence and Communication Technology,pp.718 -723,2016.

[24] S.Barman(Mandal),M.Roy,S.Biswas,S.Saha,"Prediction Of Cancer Cell Using Digital Signal Processing", International Journal of Engineering,pp. 91-95, 2011.
[25] Peng Qiu, Wang Z.Jane, and K.J. Ray Liu, "Genomic Processing of Cancer Classification and Prediction". IEEE Signal Processing Magazine, January 2007.
[26] L. Ravichandran, A. Papandreou-Suppappola, A. Spanias, Z. Lacroix, and C. Legendre, "Waveform Mapping and Time-Frequency Processing of DNA and Protein Sequences," IEEE Trans. Signal Processing, vol. 59, no. 9, pp. 4210-4224, Sept. 2011.
[27] S. Deng, Z. Chen, G. Ding, and Y. Li, "Prediction of Protein Coding Regions by Combining Fourier and Wavelet Transform," Proc. Third Int'l Congress on Image and Signal Processing (CISP), vol. 9, pp. 4113-4117, Oct. 2010.
[28] D. Gabor, "Theory of Communication," IEEE Radio Comm. Eng. J., vol. 93, no. 26, pp. 429-441, Nov. 1946.