Wavelet Analysis in Prediction and Identification of Cancerous Genes

[1]T.Thillai Gayathri., [2]Dr.S.Allin Christe.,[1]PG Scholar,[2]Associate Professor., ECE Department, PSG College of Technology, Coimbatore.

Abstract— Cancer is one of the biggest causes for mortality worldwide. In earlier days, the cancer detection was the most important and difficult task too. The classical methods for cancer detection have some disadvantages such as more system complexity, increased time consumption and also high cost. Therefore, researchers developed an idea of applying signal processing concepts in Deoxy Ribo Nucleic Acid (DNA) sequence analysis. Digital Signal Processing (DSP) is an important technique used in various analysis of DNA sequences, like exon and intron region identification, detection of gene expression and also predict the abnormalities in the exon region. DSP solves this task with less complexity and more accuracy. Nowadays, among all DSP concepts, Wavelet Transform (WT) plays an important role for analyzing DNA sequences with wide range of applications. The main objective of this paper is to predict and identify the cancerous genes using wavelet technique. The implementation is done by using MATLAB R2014a with bioinformatics toolbox. The databases are collected from NCBI website and tested for various normal and abnormal DNA sequences of Homosapien Chromosomes.

Index Terms— Digital Signal Processing (DSP), Deoxy Ribo Nucleic Acid (DNA), Discrete Wavelet Transform(DWT), Genomic Signal Processing(GSP), Electron Ion Interaction Potential(EIIP), exons, introns.

1 INTRODUCTION

According to the reports [1] available in the research of medical field, it has been understood that one of the main reason for cancer disease is genetic abnormality. Genome Signal Processing (GSP) is the fast developing area of research in detecting cancer genome. The most important necessity in this analysis is to predict the protein coding region and find the abnormalities present in the exon region (protein coding region). For analyzing DNA sequences, there are so many methods available [2]. Among them, DSP techniques have better accuracy than traditional methods. The various applications of DSP play a vital role in the analysis of genomic sequence. First, the DNA sequences are converted into numeric representations. This is called mapping technique. There are various types of representations and mapping techniques available [3]. The DNA sequences should be converted into numeric sequences by choosing any one of the mapping technique. After converting into numerical representations, DSP concepts are applied for analysis. Some DSP techniques like Discrete Fourier Transform (DFT), digital filtering, Discrete Wavelet Transform (DWT), Parametric modeling, and entropy. Specifically, Wavelet transforms have become an important mathematical analysis tool and providing a new unifying perspective on problems of cancer genome research due to its significance and advantages [4,6].

Anastassiou [5] describes the differentiation between the protein coding region and non coding region (exon and intron) by applying DSP techniques into numerical values. The protein coding region exhibits 'period-3 behavior' [3,4] whereas non-coding region does not exhibit. This is also known as 'periodicity' property.

Tao Meng [6] explains different applications of wavelet in analyzing biological problems using wavelet coefficients. This work discusses on how various types of numerical representations applied in cancer genome research. They conclude that the concepts of signal processing are best way to analyze the biological signals. The traditional Fourier transform is best suited only for stationary signals and it reveals only global periodicity. On the contrast, wavelets provide multiscale representation of the signals. The main idea behind the wavelet transform is to decompose the signal into vectors of coefficients. These coefficients contain the information of sequence characteristics at coarse and fine scales. The Global feature is available at the coarse scales of coefficients and the fine scales contain local details.

Achuthsankar S. Nair et al.[7] described a coding measure scheme employing Electron-Ion Interaction Pseudopotential (EIIP). In this paper, they have achieved the reduction of computational overhead by 75% through this representation. By substituting the EIIP value of the nucleotides in DNA sequence, a single EIIP indicator sequence is formed. This representation makes it is easy to discriminate the exon and the intron areas of the whole genome. They have achieved better results when compared to the previous results. In addition to that, EIIP exhibits the physiochemical property and plays a predominant role in the formation of exon region (protein coding region) of genomes. This fact implies that this method will stimulate a lot of researches in this area.

The Organization of the paper is as follows, in section II concepts of signal processing and molecular biology is reviewed. Section III deals with the layout and methodology of the proposed algorithm. Section IV deals with experimental results demonstrating the performance of the proposed methodology. And finally, conclusion of this paper is discussed in Section V.

2 Concepts Of Molecular biology And Signal Processing

2.1 Human Genome and DNA

The cell is the most primitive unit of all human beings. In the human cell, the central most part is called nucleus and it contains the important chemical called DNA. The DNA is responsible for making up chromosomes for transforming the genetic instruction or information from parent cell to offspring at the time of reproduction. DNA's major responsibility function is to provide information for synthesis of proteins [8]. In 1953,Watson and Crick discovered the double helix structure called DNA. It consists of two complementary strands of sugar phosphate group with the attachment of nucleotide bases. A nucleotide is defined as numerous linked and smaller components in a single strand of DNA. The nucleotide bases are of four types-Adenine(A), Guanine(G), Cytosine(C), Thymine(T). It consists of three components, Phosphate group, Nitrogen base and pentode sugar. Purines (A and T) and Pyrimidines (G and C) are the two types of nitrogen bases in DNA. The strands of DNA are always complementary in nature (i.e.) the nucleotide base A and T are always paired and G and C are always paired. Between Purine and Pyrimidine, the strands of DNA are held together by the bond which is made of hydrogen. The structure of DNA is shown in Fig(1).

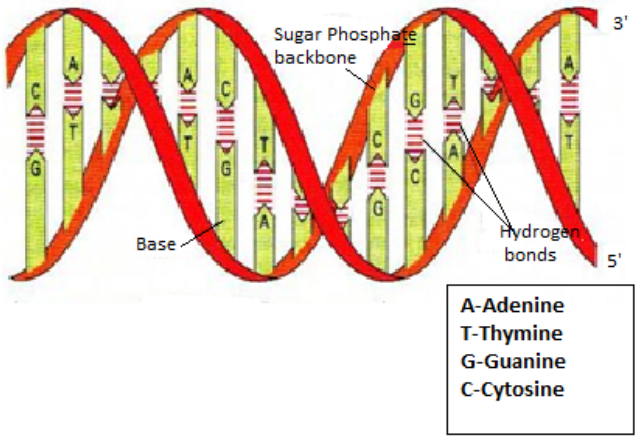


Fig.1. Structure of DNA

Cancer is mainly caused by changes (alteration) that occur in the DNA. The changes occurred in the genes are called mutations. In the human body, when the DNA within the cells gets damaged, the amount of instruction on the genes become changed or destroyed. In the cancer genome, there are four types of mutations are possible [6]. They 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. The various types of mutation are shown in Fig (2).

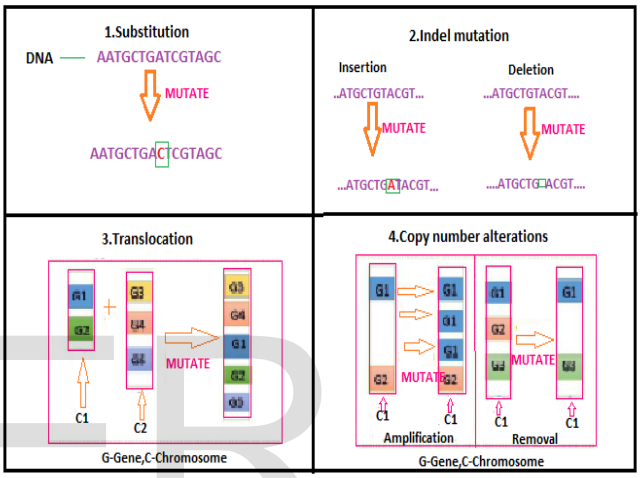


Fig.2. Classification of Mutation

2.2 Signal Processing Concepts

There are various ways for predicting the cancerous genes. Especially, DSP [10] plays a predominant role in identifying the cancerous genes. The classical Fourier Transform (FT) is one among the way to measure the composition of signal in frequency and frequency content of the signal. But in FT, the information of both time and frequency cannot be seen simultaneously. In comparison with FT technique, wavelet transform plays a predominant role because of their properties. Some of the properties are as follows,

- Wavelet transform is more useful for analyzing stationary and non-stationary signals compared to Fourier transform.
- Fourier transform is more concentrated in localization of frequency domain while wavelet transforms are localized both in time and frequency domain.
- The base functions of Fourier transform can be scaled but the wavelet transform can both be scaled and shifted.

The DFT is the major signal processing concept for converting the time domain into frequency domain. By evaluating the frequency component, the FT is used for reconstruction of the finite segment of analyzed sequence. The main drawback of FT is it becomes invalid for many of the samples if discontinuity of the sequence occurs. To discard the discontinuity problems, all processing must completed within a sample period. In genome analyzing point of view, the DFT is not efficient for small DNA sequences. The DWT is a mathematical tool for analyzing stationary and non-stationary signals. The main advantage of wavelet transform is not dependent of window length and it is used to analyze coding regions using scales.

3 METHODOLOGY

The proposed method involves the steps such as Data collection, Numerical representation, Wavelet analysis and result analysis. The flow of the proposed method is given in Fig.(3)

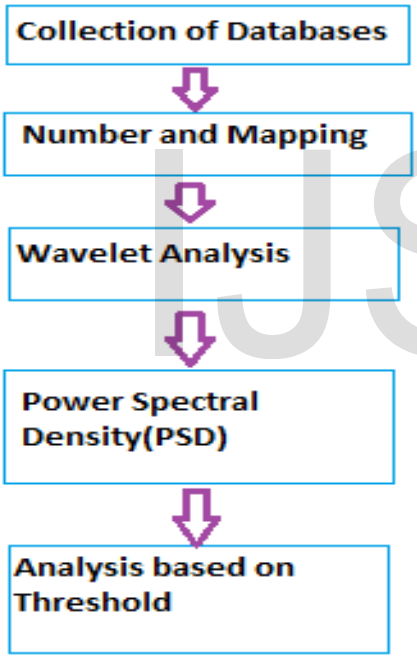


Fig.3. Flow of the Proposed Method.

3.1 Data collection

The DNA sequences are collected from the NCBI website[11]. Each sequence comprises of more than 1000 nucleotides.

3.2 Numerical Representation

Before applying any signal processing techniques, first it is necessary to convert the extracted sequence into numerical numbers using any one of the representation methods. There are so many methods available for mapping DNA sequences such as Fixed mapping technique, Voss representation, Physico chemical property based mapping, etc.[3] Here, EIIP technique is used for mapping techniques. This representation method is used because of its various advantages like DNA physiochemical property, provides biological information, improving the capability of gene discrimination and also computational overhead is reduced by 75%[12]. The EIIP value for DNA nucleotides is shown in Table 1. For example if s = [T A C G T A C G T]. By substituting the value of nucleotides from Table.1, s = [0.1335 0.1260 0.1340 0.0806 0.1335 0.1260 0.1340 0.0806 0.1335] .

TABLE.1
EIIP VALUE FOR NUCLEOTIDES

Nucleotide	EIIP Value
A	0.1260
G	0.0806
T	0.1335
C	0.1340

The sample EIIP plot for the DNA sequence TACGGCTGTC …… is shown in Fig.4.

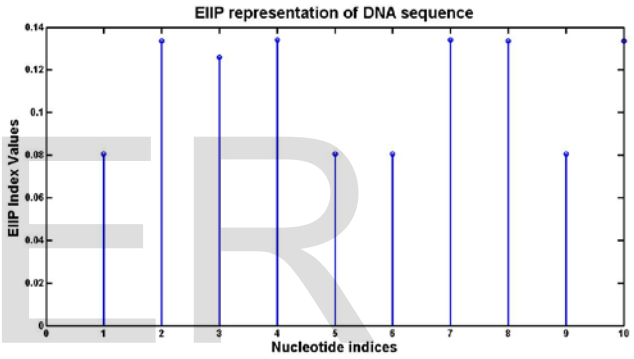


Fig.4. Sample EIIP plot for the DNA sequence GTACGGCTGTC….The x-axis represents nucleotide indices(A,T,C,G) and the y-axis represents EIIP values(0.1260,0.1335,0.1340,0.0806).

3.3 Wavelet Analysis

DWT technique is applied for feature extraction over converted DNA sequences. Most of the real world signals are non stationary signals and vary in both time and frequency domain. For biological signals, it is necessary to analyze in both time and frequency domain. Short Term Fourier Transform (STFT) [13] technique is used to analyze small portion of the signal at a time. It is termed as windowing technique. But the drawback is that it has fixed window. It leads to poor frequency resolution, when the size of the window is fixed as narrow and a wide window size leads to poor resolution of time. To overcome the problem of resolution, the wavelet transform was proposed. Wavelet transform have more advantages due to their fascinating properties like feature identification, representation of time and frequency domain, analyzing of non stationary signals, etc.

4 ALGORITHM

The algorithm of proposed method have done in two methods,

4.1 Method I: Digital spectra analysis

- The DNA sequences of non cancer cells are collected from the NCBI website.
- Symbolic DNA sequences are converted into numerical sequences using EIIP representation.
- Manually mutate the DNA sequence by various types of mutations described in section II.
- Apply Wavelet transform on the normal and mutated sequences.
- The wavelet transform have the advantage of visualizing the variations in DNA sequences at different scales due to mutation. By performing this analysis, it is easy to recognize the variations between the normal and the mutated sequence.
- The inference made by the method 1 is to identify the mutation spots occurred in the DNA sequences as shown in fig.5.

4.2 Method II: Power spectral density

- The DNA sequences of both normal and cancer cells are retrieved from the NCBI website.
- DNA sequences are transformed into EIIP sequences.
- Apply DWT technique to numeric DNA sequences and power spectral density is computed.
- The discrimination between the normal and cancer cell is determined from variations in power spectral density.
- The mean amplitude and standard deviation of a signal are computed from each spectral plot for classification.
- The coefficient of variation is also computed for differentiation of normal and cancer cells.

5 Results And Analysis

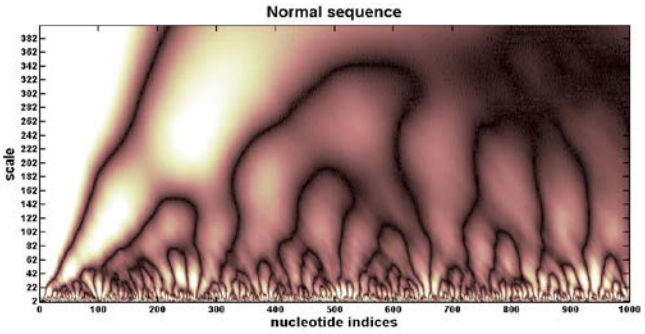
For cancer detection, the methodology developed in this paper is very much useful compared to some biological experiments. Due to low computational complexity, DWT technique is used for the identification of exon region. Several DNA sequences for both normal and cancer cells with various accession numbers is collected from NCBI website as shown in Table.2.

In the proposed algorithm, method I 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 as given in Fig.5. (ii) and (iii). Method II involves the discrimination between the normal and the cancer cells is identified by the spikes in the power spectrum plot. The spikes occurs only in the cancer cell whereas absent in the normal cell. In addition to that, the cancer cells are identified by calculating the parameters like mean amplitude, standard deviation, coefficient of variation. The ratio of mean amplitude to standard deviation is less than one which indicates cancer cells and the more than one indicates normal cell[12]. The coefficient of variation with <100% indicates normal cell and more than 100% indicates cancer cell.

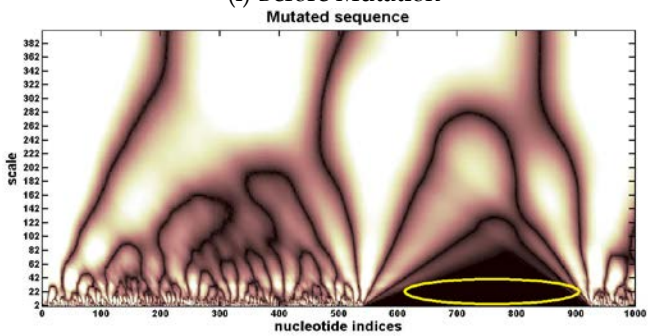
TABLE.2
NORMAL AND CANCER DNA SEQUENCES WITH ITS ACCESSION NUMBERS

S.No	Cell Types	Accession Numbers	Gene Name	Relative Position	Length of Exon
1	Non Cancer Cells	AF186616	HBB	992:1177	186
2		AF186613.1	HBB	987:1172	186
3		AF083883	HBB	1237:1425	189
4		AF186608	HBB	989:1175	186
5		AF186607.1	HBB	988:1173	186
6		AF186614	HBB	988:1173	186
7		AF186611	HBB	987:1172	186
8		AF348448	HBB	139:324	186
1	Cancer Cells	NM_004333.4	BRAF	62:2362	2301
2		EE178466	TP53	26:271	246
3		NM_007294.3	BRCA1	233:5824	5592
4		NM_000595	LTA	162:781	620
5		NM_004103.4	PTK2B	254:3283	3030
6		NM_000142.4	FGFR3	257:2677	2421
7		AF284036	KLF6	4346:5238	892
8		NM_025225.2	PNPLA3	174:1619	1446

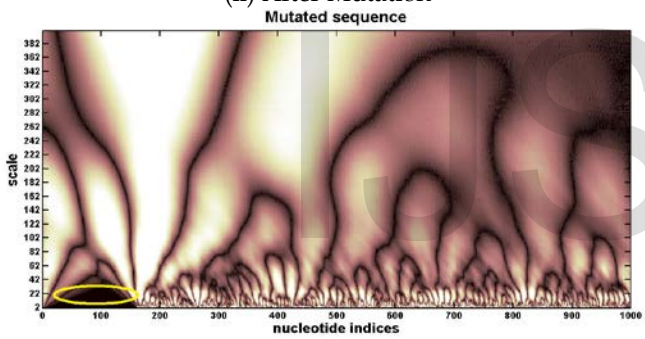
5.1 Result Analysis for Mutated and Non Mutated Sequences



(i) Before Mutation



(ii) After Mutation



(iii) After Mutation

Fig.5. Wavelet Transform plot for sequence AF186607.1. The x-axis and y-axis represent the nucleotides indices and the scale numbers, respectively. The yellow color highlights the mutated region.(ii)Substitution mutation.(substituting the nucleotides at the end of the sequence)(iii)Insertion mutation(Inserting the nucleotide at the initial level without changing any nucleotide pair).[Refer section II for details].

5.2 Result Analysis for Normal and Cancer Sequences

CANCER CELLS

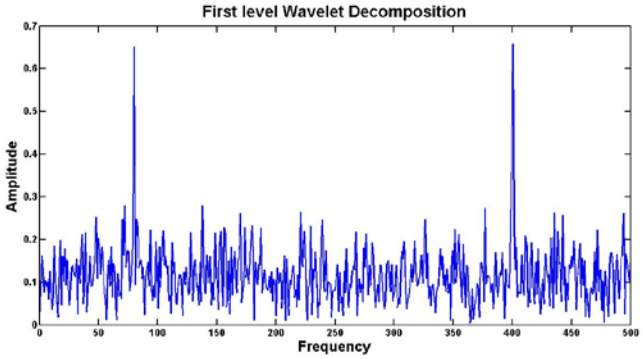


Fig. 6. Plot of Power Spectrum for Cancer Cell of Accession no. NM_004333.4

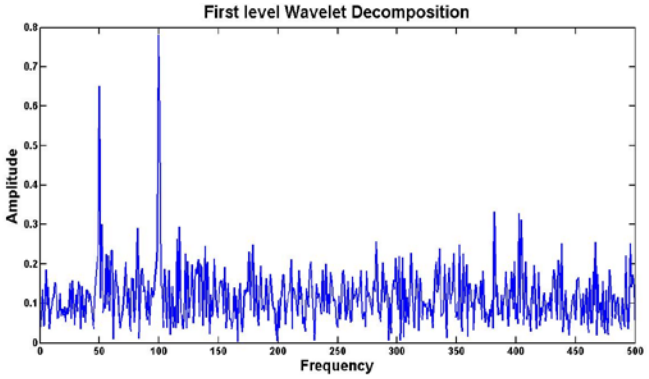


Fig. 7. Plot of Power Spectrum for Cancer Cell of Accession no. NM_000595

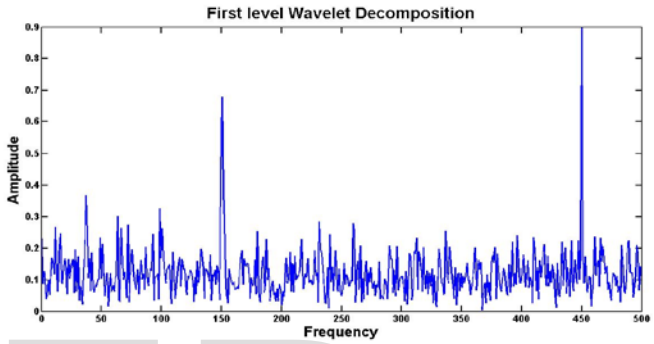


Fig. 8. Plot of Power Spectrum for Cancer Cell of Accession no. AF284036

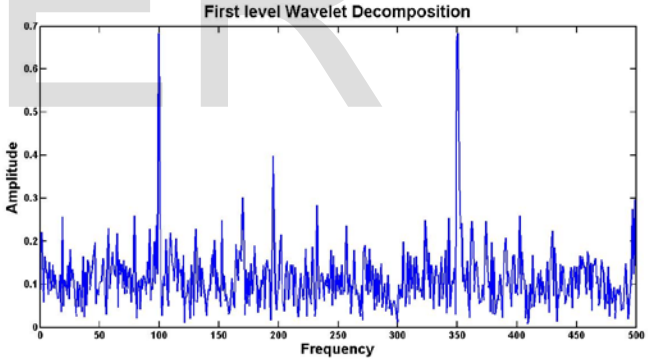


Fig. 9. Plot of Power Spectrum for Cancer Cell of Accession no. AF348515.1

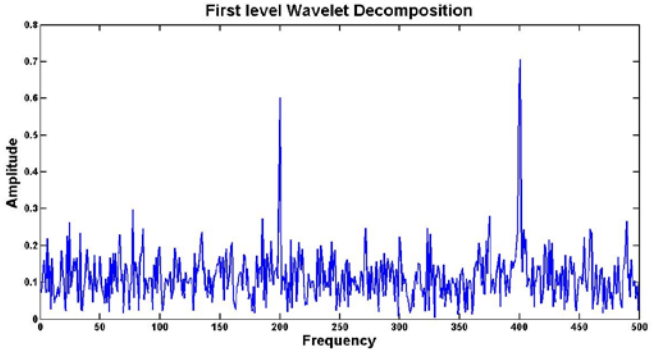


Fig. 10. Plot of Power Spectrum for Cancer Cell of Accession no. NM_005732.3

NORMAL CELLS

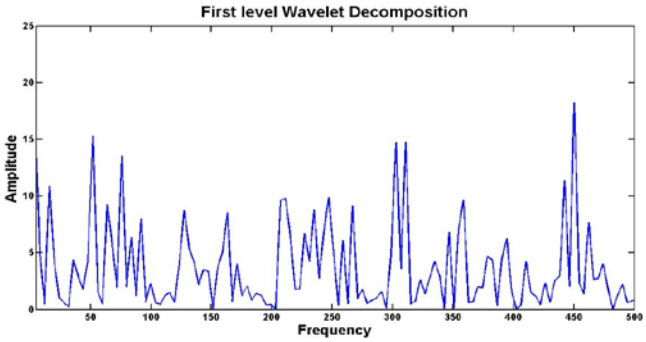


Fig. 11. Plot of Power Spectrum for Normal Cell of Accession no. AF083883

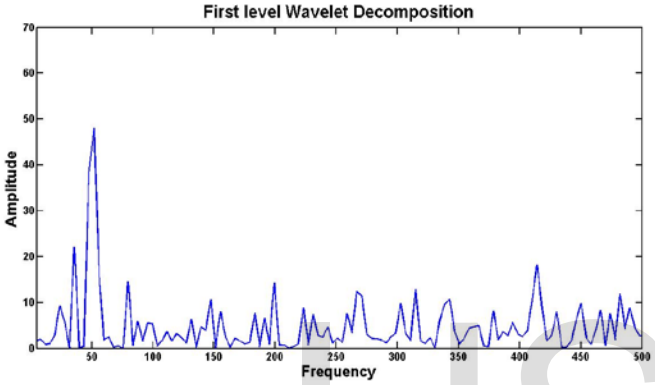


Fig. 12. Plot of Power Spectrum for Normal Cell of Accession no. AF186613.1

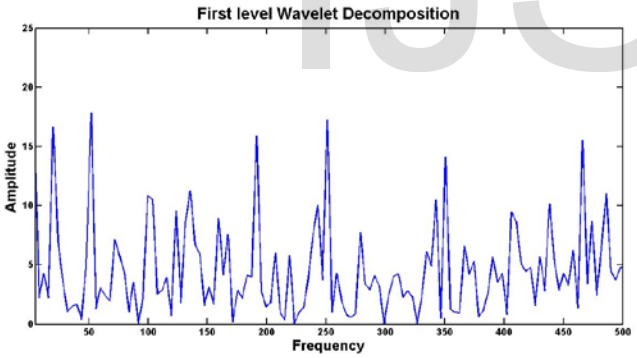


Fig. 13. Plot of Power Spectrum for Normal Cell of Accession no. AF186608

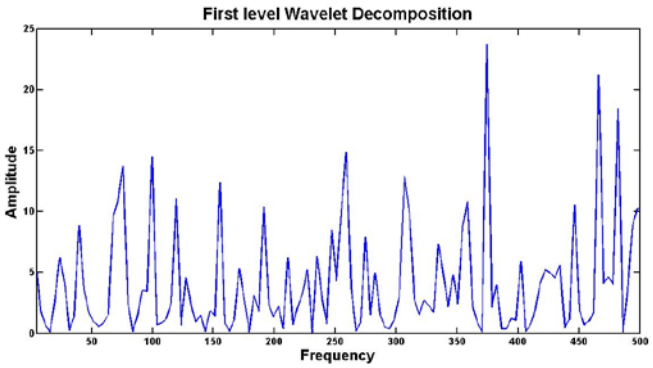


Fig. 14. Plot of Power Spectrum for Normal Cell of Accession no. AF348448

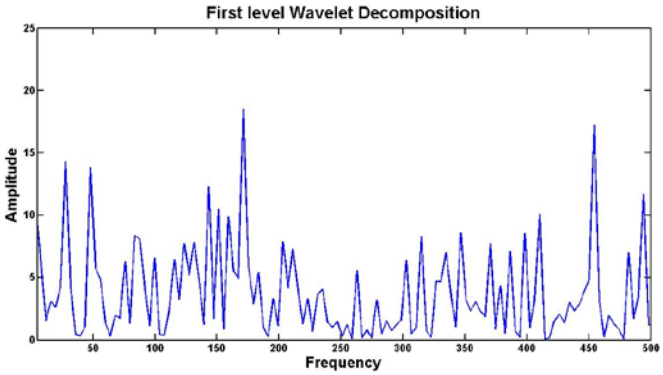


Fig. 15. Plot of Power Spectrum for Normal Cell of Accession no. AF186611

TABLE.3
ANALYSIS OF NORMAL AND CANCER CELLS

S.No	Cell Types	Accession Numbers	Mean amplitude (X)	Standard deviation (Z)	Ratio (X/Z)	Coefficient of Variation (%)
1	Non Cancer Cells	AF083883	0.0534	0.0382	1.3979	71.53
2		AF186613.1	0.0397	0.0292	1.3595	73.55
3		AF186608	0.0487	0.0321	1.5171	65.91
4		AF348448	0.0597	0.0432	1.3819	72.36
5		AF186611	0.0621	0.0547	1.1352	88.08
1	Cancer Cells	NM_004333.4	0.5678	0.9873	0.5751	173.88
2		NM_000595	0.2578	0.3526	0.7321	136.77
3		AF284036	0.3784	0.4723	0.8011	124.81
4		AF348515.1	0.0486	0.0531	0.9152	109.08
5		NM_005732.3	0.3652	0.4875	0.7491	133.48

Table [3] shows the parameters such as mean amplitude, Standard Deviation and the Coefficient of Variation for both normal and cancer cells. The inference obtained from the above table is that ratio of mean amplitude to standard deviation is less than 1 for cancer cells and more than 1 for normal cells.[12] The coefficient of Variation is more than 100% for cancer cells and less than 100% for normal cells. It is depicted via bar chart as shown in fig.16,17.

5.3 Analysis of Normal Cells and Cancer Cells

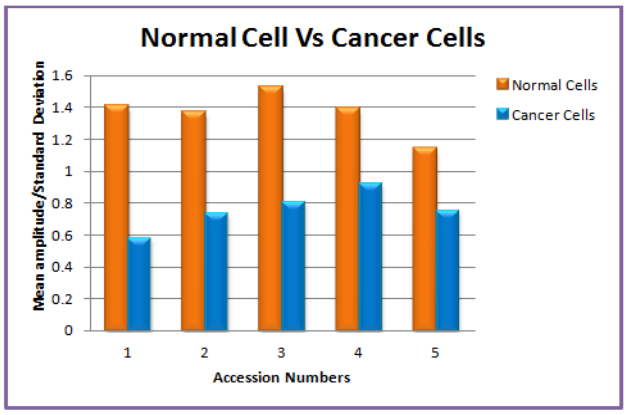


Fig.16. Ratio of Mean amplitude to standard deviation

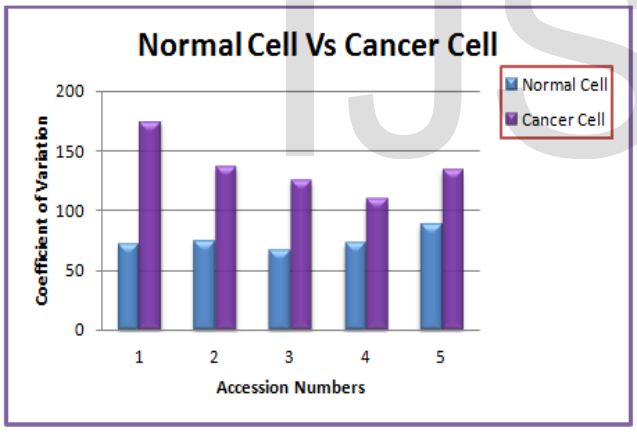


Fig.17. Plot for Coefficient of Variation

6 Conclusion

In the present work, efficient approaches have been developed for analyzing the DNA sequences. The numerical conversion technique of DNA sequences is the most important method and the conversion technique called EIIP (Electron Ion Interaction Potential) is used here. EIIP technique reduces computational overhead by 75% when compared to other representation techniques. First method involves the detection of mutational spots by applying the wavelet transform to the mutated and non-mutated sequences. Some types of mutation doesn't cause any severe effects but some induce drastic change in the DNA sequences which causes severe diseases like cancer. The effectiveness of this approach has been brought out in identifying the region of mutation in DNA sequences. The second method briefly demonstrates the discrimination between the normal and the cancer cells. In this method, the power spectral density has to be plotted for both normal and the cancer genes. The protein coding region (exon) in the power spectral plot clearly indicates the cancer caused region. Hence this micro level research has great scope in diagnosis of cancer in future.

Finally, on threshold basis it is very easy to predict the cancer genes by computing the parameters such as mean amplitude, standard deviation and the coefficient of variation. Instead of biological experiments, one can easily recognize the cancer cells with the help of signal processing concepts. In addition, it consumes less time and cost effective compared with biological concepts in cancer prediction.

References

[1] Professors D.C. Wertz, J.C. Fletcher, K. Berg, "Review of Ethical Issues in Medical Genetics" ,WORLD HEALTH ORGANIZATION(WHO), 2003

[2] Peng Qiu, Z. Jane Wang, and K.J. Ray Liu, "Genomic Processing for Cancer Classification and Prediction", IEEE Signal Processing Magazine, pp.100-110, January 2007".

[3] 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.

[4] Jianchang Ning, Charles N. Moore, James C. Nelson," Wavelet Analysis of Nucleotide Genomic Sequences",pp.1-6.

[5] D. Anastassiou, "Genomic Signal Processing," IEEE Signal Processing Magazine, vol. 18, no. 4, pp. 8-20, July 2001.

[6] Tao Meng, Ahmed T. Soliman, Mei-Ling Shyu, Yimin Yang, Shu-Ching Chen,S.S. Iyengar, John S. Yordy, and Puneeth Iyengar , "Wavelet analysis in current cancer Genome Research: A Survey", IEEE /ACM transactions on computational biology and bioinformatics, vol. 10, pp. 567-570, 2013.

[7] S. Nair, S. P. Sreenadhan, "A coding measure scheme employing electron-ion interaction pseudopotential (EIIP)," Bioinformation, vol. 1, pp. 197-202, 2006.

[8] C.Yin, SS. Yau, "Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence" Journal of Theoretical Biology , vol. 247, pp. 687–694, 10 April 2007.

[9] A. Marin, Oliver et al. "On the Origin of the Periodicity of Three in Protein Coding DNA Sequences," Journal of Theoretical Biology, vol.167, pp. 413- 414, 1994.

[10] 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.

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

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

[13] J.B. Allen and L.R. Rabiner , "A Unified Approach to Short-Time Fourier Analysis and Synthesis," Proc. IEEE, vol. 65, no. 11,pp. 1558-1564, Nov. 1977.

[14] R.Jindal, B.Banerji, Deepa Grover," Prediction and Identification of Cancerous Cells using Genomic Signal Processing",Vol 5,pp.14-23,2015.

[15] D. Gabor, "Theory of Communication," IEEE Radio Comm. Eng. J.,vol. 93, no. 26, pp. 429-441, Nov. 1946.

[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] E. S. Samundeeswari,P. K. Saranya," Computational Techniques in Breast Cancer Diagnosis and Prognosis: A Review ",International Journal of Advanced Research , Vol. 3,pp. 770 – 775,2015.

[18] Hariprasad, S.A.,Saneesh, Cleatus, Anjali Chitaranjan,Ashwini Datta N.,Monisha M. Ganesh, "Novel Approach On Cancer Detection", Proceedings Of Asar International Conference,pp. 60-63 14th May-2014.

[19] Peng Qiu, Z. Jane Wang, and K.J. Ray Liu,"Genomic Processing for Cancer Classification and Prediction", IEEE Signal Processing Magazine , pp. 100-110 January 2007.

[20] Lakshminarayan Ravichandran, Antonia Papandreou-Suppappola, Andreas Spanias, Zoé Lacroix, and Christophe Legendre ,"Waveform Mapping and Time-Frequency Processing of DNA and Protein Sequences", IEEE Transactions on Signal Processing, Vol. 59,pp. 4210-4224 ,September 2011.