THE STATISTICAL APPROACH AND OVERVIEW IN DETECTION OF CANCER CELLS BASED ON FFT AND DWT EMPLOYING GENOMICS SIGNAL PROCESSING TECHNIQUES ON DNA

Ghanshayamkumar.B. Rathod, Vatsal Shah, Neel Macwan, Sakhiya Deep Jiteshkumar, Navadiya Harshkumar Ashvinbhai

•

Birla Vishvakarma Mahavidyalaya
ghanshyam.rathod@bvmengineering.ac.in, shahvatshubh@gmail.com, neelmac094@gmail.com
sakhiyadeep007@gmail.com, navadiyaharsh111@gmail.com

Abstract

Cancer comprises a group of diseases that causes abnormal cell growth in the human body. Lakhs of people suffer from these diseases and ultimately they died due to cancer. So it is necessary to detect these diseases in an early stage. Genomics Signal Processing deals with advance research in genetics. So by applying various GSP techniques, it becomes easier to predict one of the most dangerous diseases Cancer. In this paper, we have represented the binary mapping of the raw genomic data to convert into digital data and on applying the Fast Fourier algorithm as well as the Discrete Wavelet Transform in our algorithm to predict uncertainty that is present in the coding region of DNA of the gene of Cancer cell. The purpose of this research is to provide an accurate prediction of cancer to the cancer researcher's so that the life efficiency of any cancer patient increases. We have implemented the algorithm on Matlab 2015a which consist of signal processing kit. The proposed algorithm is applying on several DNA sequences present in normal genes as well as cancer genes of Homosapiens chromosomes which is available on the National Center of Biotechnology Information (NCBI) database.

Keywords: Fast Fourier Transform; Genomic Signal Processing; DNA; Cause of Cancer; Discrete Wavelet transform; Binary Mapping; NCBI Database.

I. Introduction

Fast Fourier Transform is an algorithm that computes N Point Discrete Fourier Transform of the discrete sequences. The Discrete Fourier Transform converts the discrete sequence into frequency domain. This method is useful in many fields but computation of these method by using the equation directly becomes slow. So these transform computes the DFT very faster. As it factorizes the DFT matrix into the product of zero factors. As a result, it manages to reduce the complexity of computing the discrete fourier transform. The Fast Fourier Transform are much more accurate then discrete fourier transform. The equation of Fast Fourier Transform is:

$$x(k)=\sum_{n=0}^{(N/2)-1}[x(n)+(-1)^{k}x(n+\frac{N}{2})]W_{N}^{kn} \tag{1}$$

where $W_N^{kn} = e^{j2\pi/N}$ is twiddle factor. N stands number of point in FFT.Fast Fourier Transform is used in various application like recording, sampling of the data, Fast algorithm for discrete sine & cosine transform, Fast chebyshev approximation, solving difference equations & computation of the isotopic distribution. Genomic Signal Processing can be defined as the analysis, processing, and use of genomic signals to gain biological knowledge, and the translation of that knowledge into the system-based application that can be used to diagnose and treat genetic disease.[9] The purpose of the genomic signal processing is to combine the principles and techniques of signal processing with the understanding of the genomics to detect, classify, control, statistical and dynamical modeling of the gene networks. It is a fundamental discipline that brings the model-based analysis of the gene. Application of Genomic Signal Processing is tissue classification &discovery of the signaling pathway both based on the expressed macromolecule phenotype of the cell. The accomplishment of these applications is done by signal processing techniques such as Fast Fourier Transform, Discrete Wavelet Transform & Discrete Fourier Transform. DNA was first identified by Francis Crick and James Watson at the Cavendish Laboratory at Cambridge University. It is a double-helical structure comprises of four nucleic acid Adenine (A), Cytosine (C), Guanine (G), Thymine(T) a deoxyribose & phosphate group. The Adenine is connected with Thymine by hydrogen bond & The Guanine is connected with Cytosine by a hydrogen bond. Both polynucleotides of DNA contain some biological information in it. This information is replicated when the polynucleotides are get separated. RNA is produced by DNA through the transcription process. DNA usually occurs as linear chromosomes in Homosapiens. The set of chromosomes makes up the human genome. The genome has 3 billion base pairs of DNA. The biological information that is carried by DNA occurs in a piece of sequences of DNA called genes. Transmission of the genetic information occurs via complementary base pairing. The helical structure of the DNA is shown below in Figure 1. [3]. DNA molecules store the digital information that shapes the genetic scheme of living organisms [2]. By understanding the structure and properties of DNA one can predict the genetic diseases.

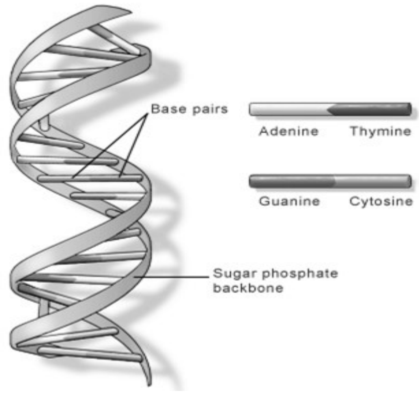

Figure.12.*Helical Structure of DNA*

Cancer is a deadly disease caused due to abnormal growth of the cells in the body and it destroys the body tissue. It begins with a genetic mutation. So because of the genetic mutation, there is unorderly growth of the cells in the body. The type of cancers is Breast Cancer, Prostate Cancer, Colon Cancer, Lung Cancer, etc. There many other types of cancer but these types of cancer are commonly found in the body. Breast cancer mostly occurs due to abnormal growth of the cells in the breast it mostly occurs in the men and rarely occurs in the women. The gene name BRCA 1 and BRCA 2 are responsible for Breast cancer. Breast cancer mostly begins with the milk-producing ducts. It is essential to predict such cancer because doctors have found that 5 to 10 percent of breast cancer genes are heritable in nature and it is passed over the family. The prostate is a walnut-size gland located between the penis and urinary bladder and the formation of the abnormal cells near the prostrate is known as prostate cancer. The genes CD82 CDH1 & CHEK2 is responsible for prostate cancer. The colon is a large intestine. It is an organ or part of the digestive system of the human body the abnormal growth of the cells in the colon causes colon cancer. The genes MSH 1and MSH 6 is responsible for colon cancer. This cancer is hereditary in nature. Lung cancer occurs due to the abnormal growth of the cells in the lungs. It mostly occurs with the person who smokes The risk of lung cancer increases with the number of cigarettes that individuals smoke if individuals quit smoking after smoking many years then the individual has reduced the chances for lung cancer development in their body. Figure 2 [4] indicates how genetic mutation occurs in a human body that causes cancer. By applying our algorithm, it becomes easy for the researcher to target cancer and to develop such devices or drugs which improve the life span of the patient. we design a mathematical model for cancer cell prediction by applying various signal processing algorithms.

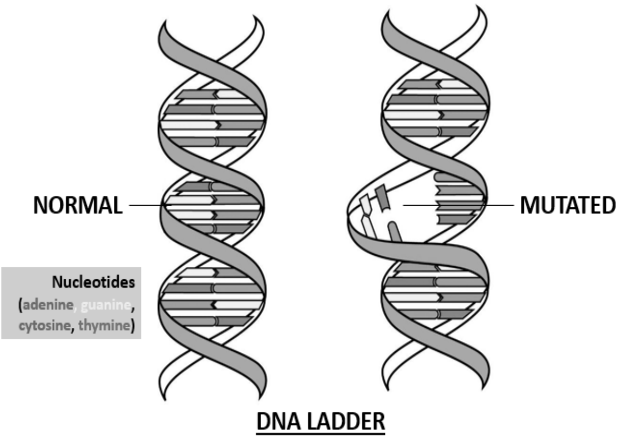

Figure.13.*Genetic Mutation in DNA Structure*

A wavelet is a mathematical technique that is used to abstract particular data from many types of data. Discrete wavelet transform means the transform which divides the signals into two orthogonal sets of the wavelets. The first discrete wavelet transform invented by Alfred Haar. If an input has 2^n numbers, the Haar wavelet transforms may be consider to pair up the input values, storing their differences and passing the sum. This process is repeats in recursion to prove the next scale which leads to 2^n -1 differences and a final sum. we have used here Haar wavelet transform the formula of the Haar wavelet transform is: The figure below shows wavelet transform [5].

$$y_n = H_n x_n \tag{2}$$

Where: H_n is a Haar Matrix

Y_n is an output

X_n is an input signal

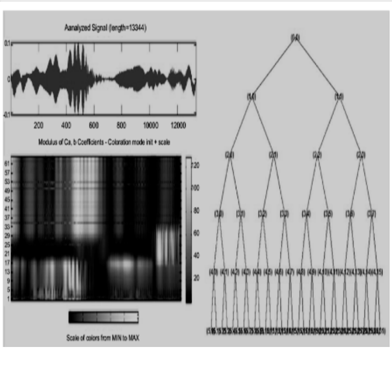

Figure.14. *Haar Wavelet Transform*

The NCBI Stands for National Center for Biotechnology Information. Which consists of information regarding genes genomic data & biomedical information. It consists of genes not limited up to humans but all organisms. It also consists of information regarding protein chains as well as gene expression, Taxonomy, Sequence Analysis. one can easily access the genomic data in order to predict the disease. The front end of the NCBI website is shown in figure 4.[6]

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

Read about NCBI resources in 2020 Nucleic Acids Research database issue

The 2020 Nucleic Acids Research

NLM announces Curation at Scale Workshop

Data curation plays a critical role in

Rapid access to 2019-nCoV (Wuhan coronavirus) data from the current public health emergency

You are here: NCBI > National Center for Biotechnology Information

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA
Policies and Guidelines | Contact

Figure. 15.*NCBI Website*

As we know that DNA consist of four chemical elements they are A(Adenine), C(Cytosine), G(Guanine), T (thymine). So binary mapping means mapping of these elements into the binary one. This is one of the most important element in our research. The table below shows the binary mapping of DNA. By using these Binary mapping concept, we have predicted the cancer cell. we have mapped the DNA sequence to its binary equivalent using Matlab 2015a.

Table1:*Binary Mapping of DNA*

DNA Components	Binary Equivalent	Decimal Value
Adenine	**00**	**0**
Cytosine	**01**	**1**
Guanine	**10**	**2**
Thymine	**11**	**3**

II. Methodology of the Research

The method that we have applied in the prediction of the cancer cell is as follows:

- First, take a DNA sequence of the Cancer gene as well as the Normal gene from the database available on the NCBI website.
- Then apply the Binary mapping into the DNA sequence.
- Thereafter find the Fast Fourier Transform of the converted binary sequence for Analysis.
- Find out the Discrete Wavelet Transform of the resultant binary coded DNA Sequence for Statistical analysis.
- Determine the ratio of the change in mean amplitude (ΔX) by Standard Deviation (S).
- If $\frac{\Delta X}{S}$ is greater than 0.5 then the Predicted Genesis of Normal cell.
- If $\frac{\Delta X}{S}$ is lesser then 0.5 then the Predicted Genesis of cancerous cell.

III. Flowchart of the Research

The flow chart below indicates the Algorithm of the Research.

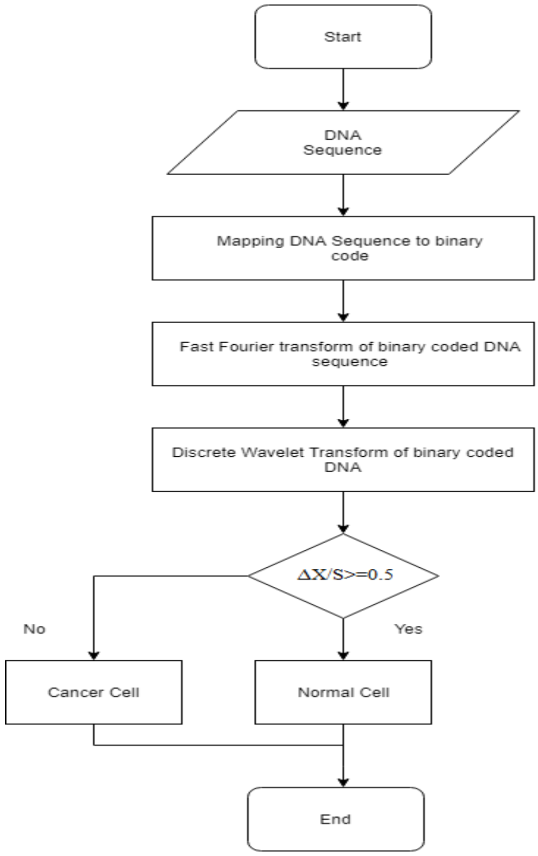
Start
DNA Sequence
Mapping DNA Sequence to binary code
Fast Fourier transform of binary coded DNA sequence
Discrete Wavelet Transform of binary coded DNA
ΔX/S>=0.5
No
Yes
Cancer Cell
Normal Cell
End

Figure. 16. Flowchart of the Research

IV. Fast Fourier Analysis

The Results for the Normal Cell and Cancerous Cell is shown below:

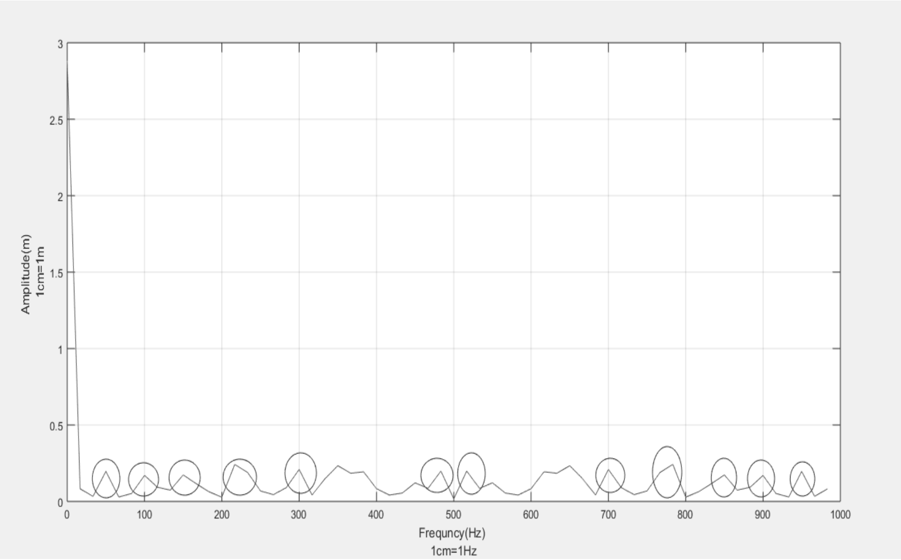

Figure .17. *BRCA1 Cancerous Cell*

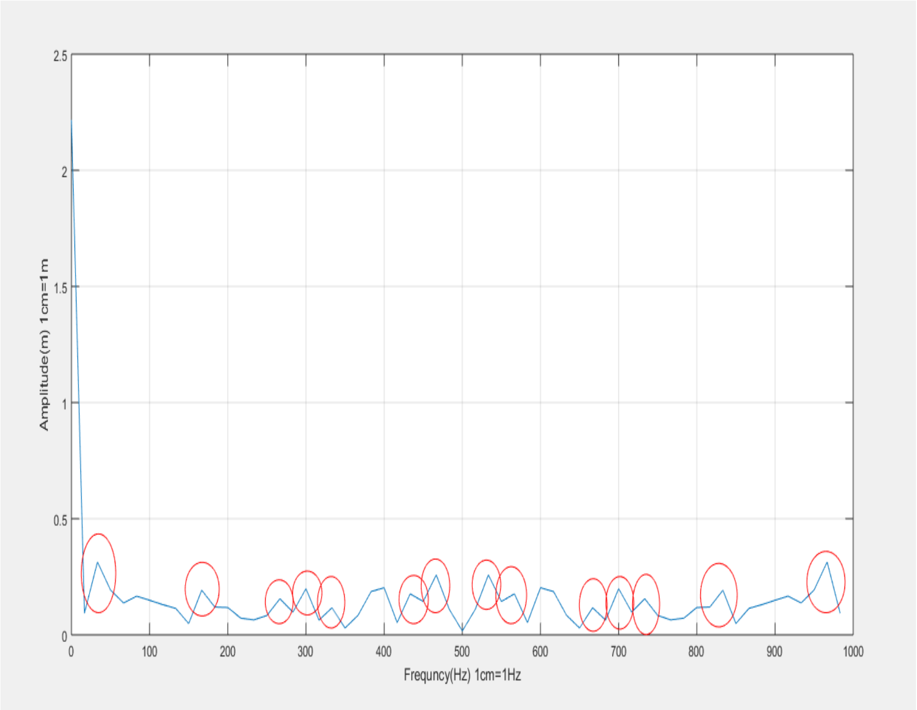

Figure. 18 . *PCA5 Cancerous Cell*

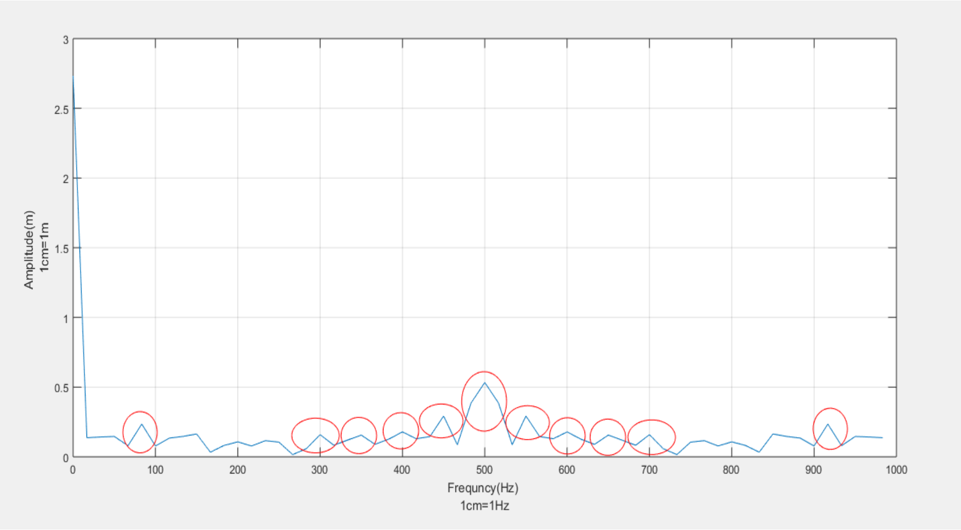

Figure .19. *TP63 Cancerous Cell*

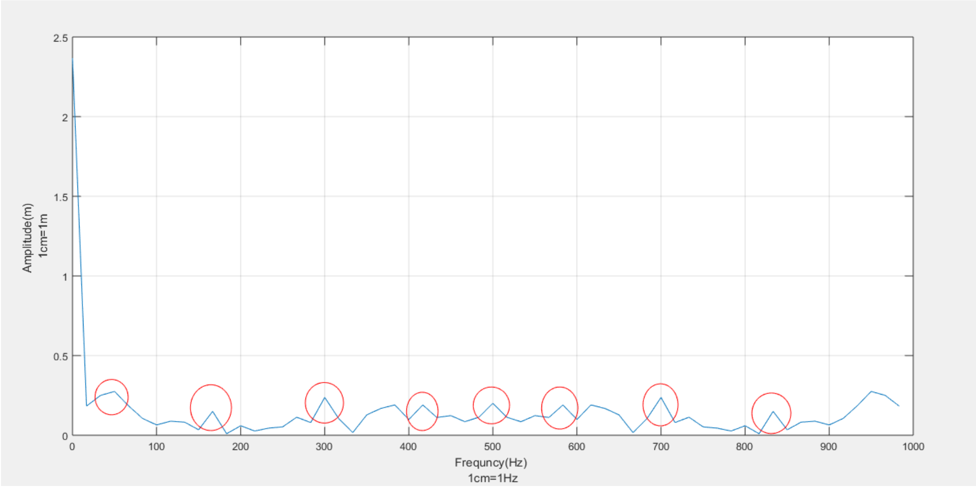

Figure. 20. *HBA2 Normal Cell*

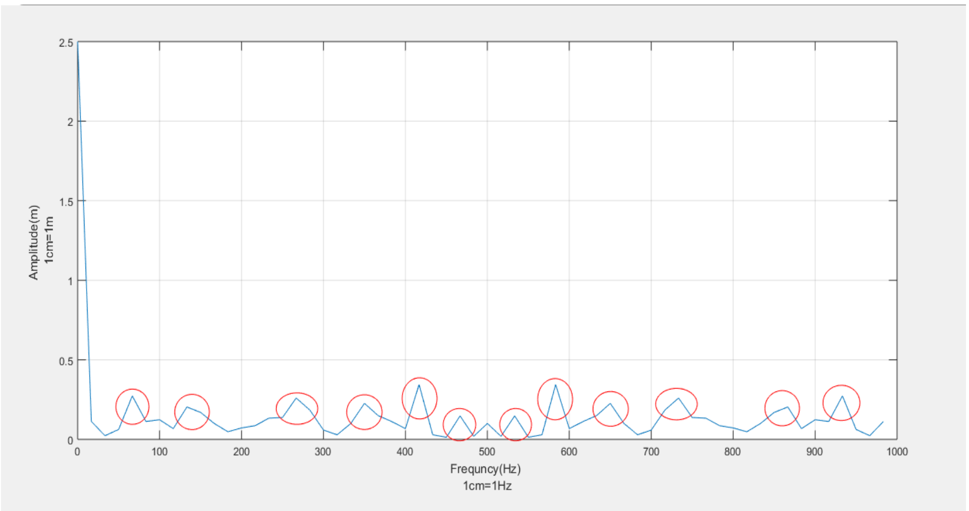

Figure .21. *HBG2 Normal Cell*

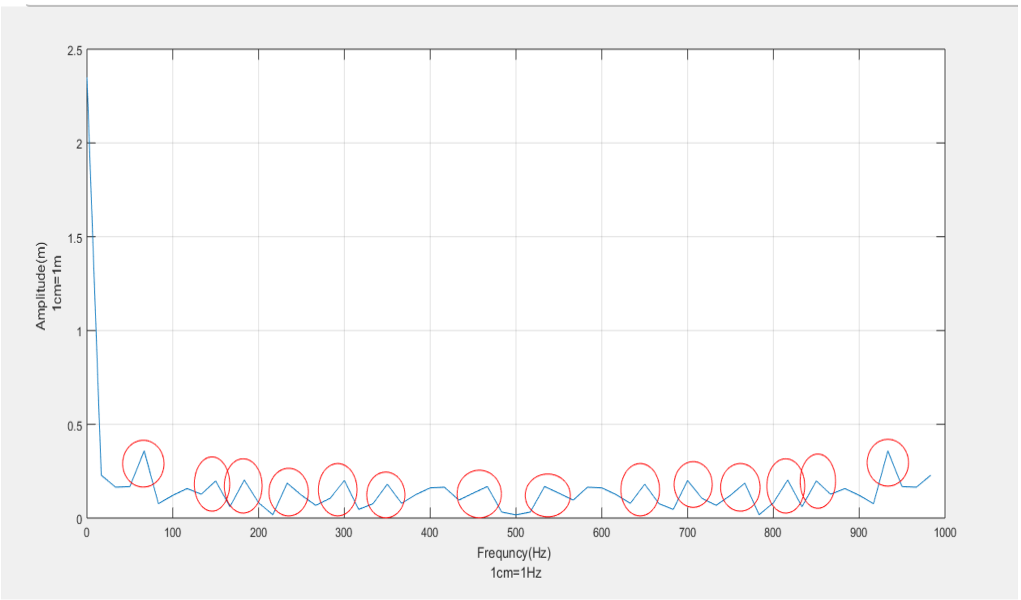

Figure .22. *Globin Normal Cell*

From these results, we have concluded that there are average 13 peaks in the waveform of the cancerous cell while there are average 11 peaks in the normal cells this means that the cancerous cells contain more noise compared to the Normal Cell.

V. Statistical Analysis

Table.2: *Genes of Cancerous Cell*

SR. No:	Name of Gene	Change in Mean Amplitude of Signal (ΔX)	Standard Deviation (S)	$\frac{\Delta X}{S}$
1	BRCA 1	0.075	0.6725	0.111
2	BRCA 2	0.241	0.5921	0.407
3	EGFR	0.125	0.6455	0.193
4	TP63	0.233	0.8703	0.2677
5	Estrogen	0.317	0.7988	0.3960
6	PCA5	0.208	0.6606	0.3148
7	GST	0.35	0.7636	0.4583

Table.3: Genes of *Normal Cell*

SR. No:	Name of Gene	Change in Mean Amplitude of Signal (ΔX)	Standard Deviation (S)	$\frac{\Delta X}{S}$
1	HBA1	0.45	0.7933	0.5644
2	HBA2	0.45	0.7811	0.57611
3	HBG1	0.9	0.922	0.97613
4	HBG2	0.917	0.7912	1.15896
5	GATA	0.917	0.7901	1.15877
6	GLOBIN	0.45	0.7973	0.5644

From the Statistical analysis we can say that if the ratio of change in mean amplitude of signal by standard deviation is greater than or equal to 0.5 then the predicted gene is of normal cell. Otherwise the predicted gene is of cancer cell table 2, 3 justify the same.

VI. Conclusion

In the presented work, an effective algorithm has been developed using MATLAB which consist of signal processing toolbox. And the most important is the mapping of the DNA sequences the mapping technique that is used here is Binary Mapping. A combination of the Fast Fourier based spectral analysis & wavelet-based methods were found to be more accurate due to their properties, such as feature extraction time-frequency domain representation, multi-resolution, scalability, de-noising and also compressing of big sample data or sequences of data. Fast Fourier analysis technique has been applied to raw genomic data for detection of the uncertainty of DNA sequences. So in the future, it has a great scope in early diagnosis of cancer as it depends on the permanent alteration in the DNA sequence of the gene. This can help in developing drugs, to drug designer. From information that one can get from genomic data one can determine how an individual will respond to particular drug and based on that design of new drugs will be done. And also without performing experiments, one can determine accurately diseases with the help of digital signal processing theory and techniques which further leads to a cost-effective experiment as it is non-invasive and require less amount of time.

References

[1] Hong-Qiang Wang, Hau-San Wong, De-Shaung Huang, Jun Shu." Extracting gene regulation information for cancer classification", Pattern Recognition,2007.

[2] Safaa M. Nadeem, Mohamed A. Eldosoky, Mai S. Mabrouk "Detecting genetic Variants-Breast cancer using different power spectrum methods", pp.147-153.

[3] "What is DNA"? -Genetics Home Reference-NIH," U.S. National library of Medicine.

[4] what-is-a-dna-mutation"? -available at dnasu.com

[5] researchgate.net/figure/Signal-processing-with-haar-wavelet-in-a-Analyse-signal-and-b-Decomposition-Tree-of_fig4_303922065

[6] ncbi.nlm.nih.gov

[7] Shilpi Chakraborty, Vinit Gupta "DWT based Cancer Identification using EIIP".2016 Second International Conference on Computational Intelligence and Communication Technology, pp.718-723.

[8] G.N. Satapathi, P. Srihari, A. Jyothi, S. Lavanya," Prediction of cancer cell using digital signal Processing". International Conference on Communication and Signal Processing, vol. 23, pp. 149-153, 3 April 2013.

[9] Worldwide Science.org