ANNALS OF FACULTY ENGINEERING HUNEDOARA
- INTERNATIONAL JOURNAL OF ENGINEERING
Tome IX (Year 2011). Fascicule 3. (ISSN 1584 - 2673)

[1.] S.BARMAN (MANDAL), [2.] M.ROY, [3] S.BISWAS, [4.] S.SAHA

PREDICTION OF CANCER CELL USING DIGITAL SIGNAL PROCESSING

[1,3,4] INSTITUTE OF RADIO PHYSICS & ELECTRONICS , UNIVERSITY OF CALCUTTA, KOLKATA-700009, INDIA
[2] THE CALCUTTA TECHNICAL SCHOOL, 110 SN BANERJEE ROAD, KOLKATA -700013, INDIA

ABSTRACT: Digital Signal Processing (DSP) applications have gained great popularity in the study of genomics in recent time. DSP can be used as a tool in the area of DNA sequence analysis, gene expression detection, identification of coding and non coding regions and also finding out abnormalities present in the coding region. DSP solves this task with great accuracy and less complexity. According to available medical research reports it has been given to understand that cancer is often caused due to genetic abnormality. In the present article, we present a DFT based approach to analyze the spectral characteristics of cancer cell and non- cancer cell and design a digital IIR low pass filter with Butterworth approximation for better prediction and identification of anomalies in cancer cells. The algorithm is tested for several databases of Homo Sapiens chromosomes available in Gene Bank which gives satisfactory results.

KEYWORDS: Genomic Signal Processing; DSP in Cancer prediction; DFT power spectrum; DNA sequence analysis; coding region

❖ INTRODUCTION

Genomic signal processing (GSP) is the engineering discipline that studies the processing of genomic signals, which are measurable events originating from DNA sequence, mRNA sequence and protein. Based upon current technology, GSP primarily deals with extracting information from gene expression measurements. The analysis, processing and use of genomic signals for gaining biological knowledge constitute the domain of GSP. Cells are the fundamental working units of every living system. All the instructions needed to direct their activity are contained within the chemical bases of a DNA chain. When a particular instruction becomes active the corresponding gene is said to be turned on or be expressed. Two major goals of functional genomics are 1) to use genomic signals to classify disease on a molecular level and 2) to screen for genes that determine specific disease and model their activity in such a way that normal and abnormal behavior can be differentiated. Over the past decade, significant discoveries have been made that provide a better understanding of genetic basis of cancer disease. It has been understood that DNA plays an important role in the study of cancer disease.

Now a days Cancer is the most common and dreaded disease that plays a leading role causing death all over the world. Almost in every family there is at least one victim of this disease. Cancer causes a significant financial burden on the health care system. Despite many advances derived from important innovations in technology during the last decades, in the field of cancer medicine, successes are still overshadowed by the tremendous morbidity and mortality incurred by this devastating disease. Cancer is caused by abnormalities in the genetic material of the transformed cell. Cancer-promoting genetic abnormalities may randomly occur through errors in DNA replication or are inherited and thus present in all cells from birth. The heritability of cancer is usually affected by complex interactions between carcinogens and the host's genome. There is a diverse classification scheme for the various genomic changes which may contribute to the generation of cancer cells. Most of these changes are mutations, or changes in the nucleotide sequence of genomic DNA. Small-scale mutations include point mutations, deletions, and insertions, which may occur in the promoter region of a gene and affect its expression or may occur in the gene's coding sequence and alter the function or stability of its protein product [10]. Most cancers come from random mutations that develop in body cells during one's lifetime - either as a mistake when cells are going through cell division or in response to injuries from environmental agents such as exposure to radiation or chemicals. Many researchers are using gene microarray and mass spectrography technology to analyze gene expression data for prediction and classification of cancer [5]. Nanotechnology is also used to develop accurate and sensitive biomedical devices for cancer genome study [3].

Since abnormality of the DNA and coding regions are related to cancer, we focus our attention to study the spectral characteristic of coding regions using digital signal processing for cancer cell and non-cancer cells. In this paper the authors have presented DFT power spectrum methods to predict abnormality present in the nucleotide levels of a coding region in the cancer cells and also design an IIR Low Pass digital filter with Butterworth Approximation for better prediction of cancer cells.

Brief Overview of DNA

DNA is the important chemical present in the nucleus of all cells. It makes up chromosomes which is responsible for passing on the genetic information from parent cell to offspring during reproduction. The major function of DNA is to provide instructions for protein synthesis [1]. After the sensational discovery of double helix structure of DNA by Watson & Crick, researchers from all fields have focused their attention in this particular field of biology keeping in view the vast information content and functional importance of DNA. A DNA is a double helix structure consisting of two complementary strands of sugar-phosphate group with bases attached to it. The sequence of nucleotide bases A, T, G and C (respectively, adenine, thymine, guanine and cytosine) provide all genetic information needed to carry out cell's activities. Nucleotides are made up of three components nitrogen base, pentode sugar and phosphate group. Two types of nitrogen bases purines and pyrimidines are present in DNA. A and G are purine bases whereas C and G are pyrimidine bases. Two strands of DNA are always complementary to each other. A always pairs with T and C always pairs with G. The DNA strands are held together mainly by hydrogen bonds between purine and pyrimidine. There are two Hydrogen bonds between A & T and three bonds between C & G. Therefore GC bond is stronger than AT bond as shown in Figure 1.

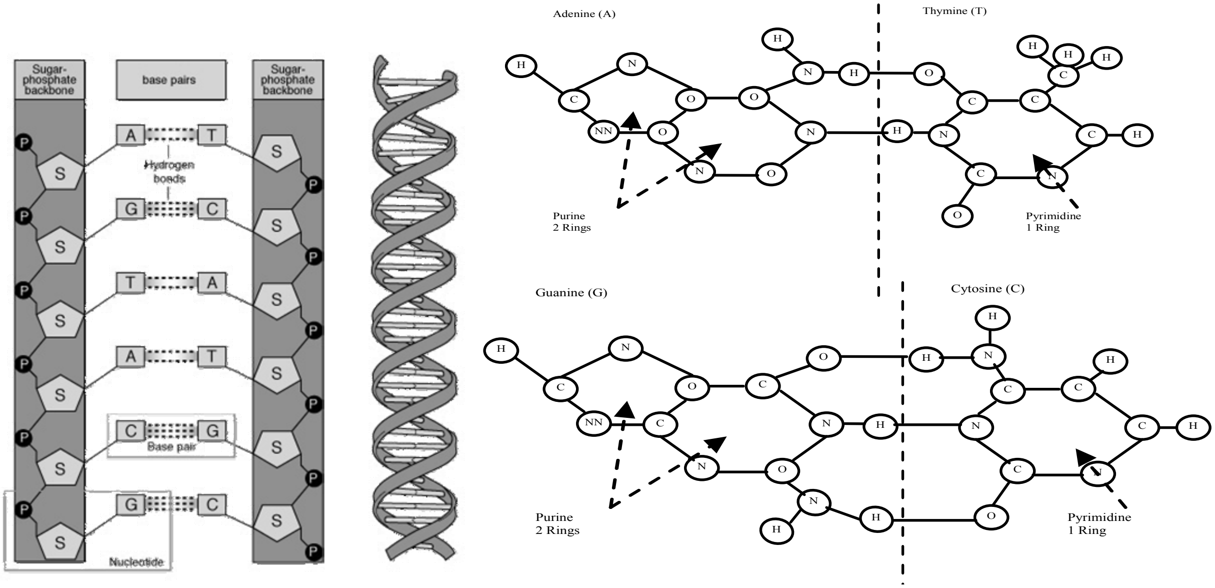


Figure 1. DNA structure and hydrogen bonding between neucleotides

Exons of a DNA sequence are the most information bearing part because only exons involve in protein synthesis [2,6,7,14]. Hence Power Spectrum of exons has been used in the study of Cancer cells in this paper. A mutation is a permanent change in the DNA. A mutation can arise spontaneously without apparent cause, or in response to radiation from UV light, exposure to certain chemicals or viruses. Some mutations involve the substitution of one base pair for another, for example inserting a G-C where there should have been a T-A pair. In other cases one or more base pairs can be added to or removed from the chain. Sometimes huge segments are altered, rearranged or misaligned. Mutations can lead to cell death, to alterations in the way a cell functions or in some cases to cancer. In this paper spectral analysis of cancer cell is studied based on abnormality present in the nucleotide level of a DNA sequence.

DSP technique for cancer prediction

A key concept in DSP is the possibility of representing the signals in the frequency domain making use of the Discrete Fourier Transform [9]. This representation leads to some important signal properties which are associated to their frequency spectrum that are not revealed in the time domain.

In case of the genomic sequences the nucleotide bases are represented mathematically by character strings of size -4 alphabet consisting of the letters A,T,G and C. The possibility of finding wide applications of DSP techniques to the analysis of genomic sequences occur only when these are appropriately converted into numerical sequences [4,11,12]. Several mapping techniques are used by researchers for their applications. A new mapping technique of conversion based on weak and strong hydrogen bonding between nucleotides bases has been used in this paper. Here instead of taking four independent indicator sequences for each nucleotide one binary sequence is generated based on weak & strong hydrogen bonding. For conversion we have taken 'A's & 'T's to be binary '0' and 'C'& 'G' as binary '1' .

For a DNA string defined as x[n] with alphabets A,T, C & G of length N . Let us define a single binary indicator sequence xs[n].

If x[n] = A T T G C A G C T , then xs[n] = 0 0 0 1 1 0 1 1 0

The DFT Xs[k] of the corresponding binary sequence is

$$X_s[k] = \sum x_s[n] e^{-j2\pi nk/N} \tag{1}$$

k=0,1,2,…….N-1 and n=0,1,2,…….N-1.

Then the Power Spectral content of the sequence is given by

$$P_s[k] = \sum |X_s[k]|^2 \tag{2}$$

The plot of Ps[k] of coding region is investigated as indicator of cancer cell or non cancer cell. Spectral characteristics of coding region is further improved by using a recursive filer with the following system function

$$\text{H(Z)} = \frac{b_0 z^N + b_1 z^{N-1} + b_2 z^{N-2} + \ldots\ldots + b_N}{Z^N + a_1 z^{N-1} + a_2 z^{N-2} + \ldots\ldots\ldots + a_N} \tag{3}$$

A digital IIR filter of Butterworth approximation have been designed to suppress the noise from the power spectrum providing better prediction of cancer cell.

❖ ALGORITHM

For a coding sequence of length N, find the Power Spectral Density.

Step 1: Scan the sequence and map the four letter alphabet into one binary sequence based on weak and strong hydrogen bonding between nucleotides.

Step 2: Obtained the Discrete Fourier Transform of the binary sequence using Eq.1.

Step 3: Estimate the PSD of the sequence obtained using Eq. 2.

Step 4: Design an IIR Low Pass filter with Butterworth approximation with following specification:

Filter order N: 10;

Pass band edge frequency:

Wp =0.3142 rad/sec;

Stop band edge frequency:

Ws =0.4147 rad/sec;

Cutoff frequency: Wc =0.3607 rad/sec;

Pass band ripple: Rp =0.5 db;

Stop band ripple: Rs =40 db;

The filter response is shown in Figure 2.

Step 5: Filter the PSD of the sequence using the above designed filter.

Step: 6: Find out mean amplitude, mean frequency for various sequences.

Step 7: Compare the PSD for cancer cell and non cancer cell

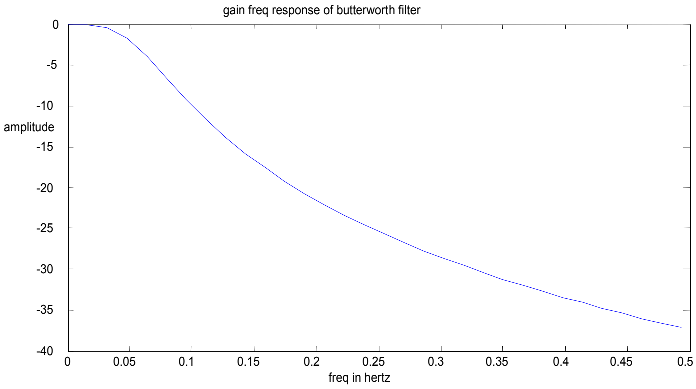


Figure 2. IIR filter Response

The algorithm is tested for various Homo sapiens cancer and non-cancer data bases available in Gene bank shown in Table 1 & Table 2 and giving satisfactory result.

Table 1. Cancer Cells

Sl. No .	Accession No.	Mean Normalized Frequency	Mean Amplitude	Ratio of Mean Amplitude and Mean Normalized Frequency
1	AF008216.1	3.786	1.976	0.52
2	AF348525.1	3.874	1.189	0.30
3	NM_016346.2	2.548	1.369	0.53
4	NM_005732.3	2.491	2.007	0.80
5	AF348515.1	3.895	0.9737	0.24
6	NM_012403.1	3.786	1.976	0.52

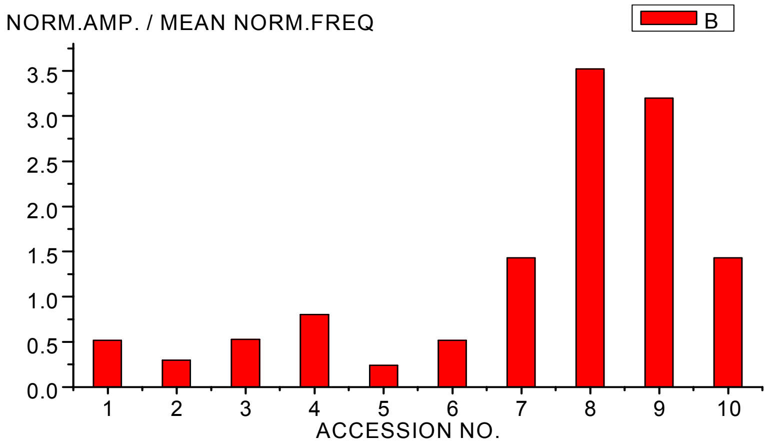
Table 2. Normal Cells

Sl. No.	Accession No.	Mean Normalized Frequency	Mean Amplitude	Ratio of Mean Amplitude and Mean Nor. Frequency
1	AF083883	3.481	5.02	1.43
2	AF186607.1	2.984	10.52	3.52
3	AF186613.1	2.984	9.565	3.2
4	AF007546	3.481	5.02	1.43

❖ RESULTS & INFERENCES

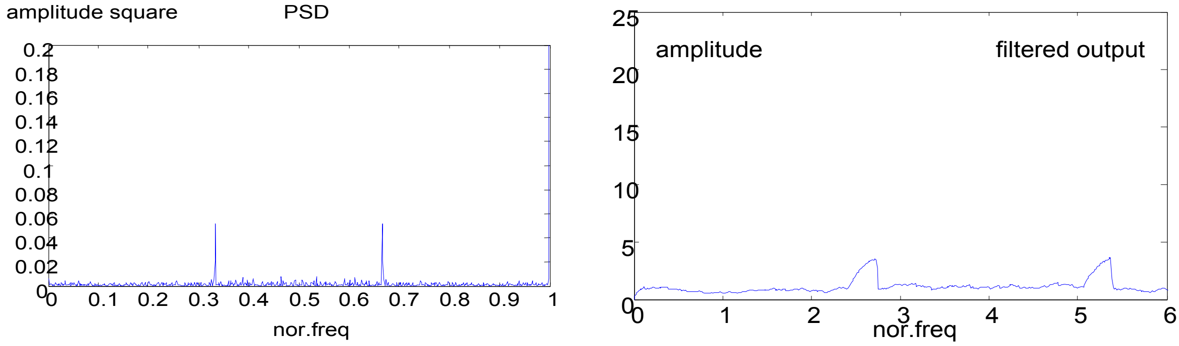
Here we have observed the DFT power spectrum plot of nucleotide bases of coding regions for various DNA sequences to identify some features for easy diagnosis of cancer disease. DFT power spectrum plot can be used as an effective and simple method for cancer prediction if DNA nucleotide sequences are available.

Several databases of the breast, ovarian, prostate cancer with different accession numbers have been taken for our sample test. Some of these data bases are shown in the Table 1.& Table 2 . From the DFT power spectrum plot depicted in Figures 4, 5, 6 & 7 of coding region, it has been observed that there is a clear distinction between cancer cell and normal cell. In most of the cases of cancer cells, spikes are generated in the power spectrum plot which is absent in normal cells. The IIR filter output gives a clear distinction of Cancer cell and normal cells. The filtered output of coding region for normal cell is more smooth, whereas cancer cells are either peaky in nature or vary randomly. It has been also observed that the filtered Power Spectrum plots for Cancer cells starts from non-zero value whereas non-cancer cell starts always from zero value of the amplitude. Table (1 & 2) summarizes the experimental results of mean amplitude, mean normalized frequency obtained in cancer cells and non cancer cells. From the bar plot shown in Figure 3. , it has been observed that ratio of mean amplitude to mean frequency is less than 1.0 for cancer cells and more than 1.0 for normal cell.



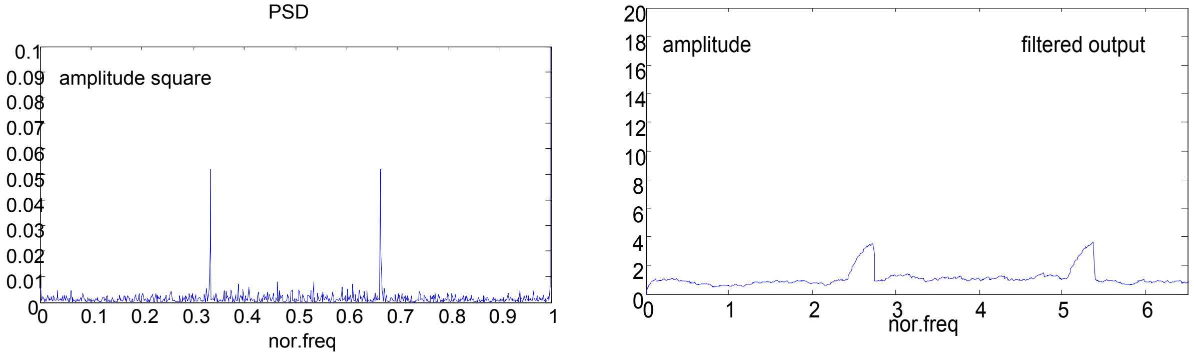
1	CANCER CELLS	AF008216.1
2		AF348525.1
3		NM_016346.2
4		NM_005732.3
5		AF348515.1
6		NM_012403.1
7	NON-CANCER CELLS	AF083883
8		AF186607.1
9		AF186613.1
10		AF007546

Figure 3. Bar Plot for Cancer cell and normal cell



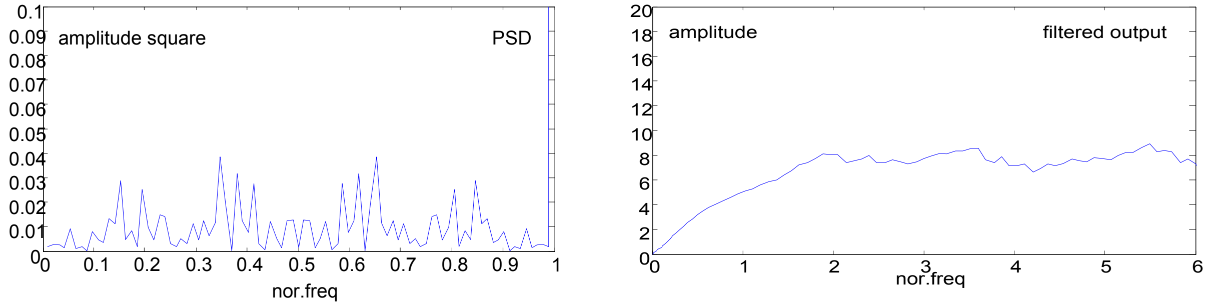
Without Filter With Filter

Figure 4. DFT Power Spectrum Plot For Cancer Cell of accession no. AF008216 and length of EXON 4453-5157 bp



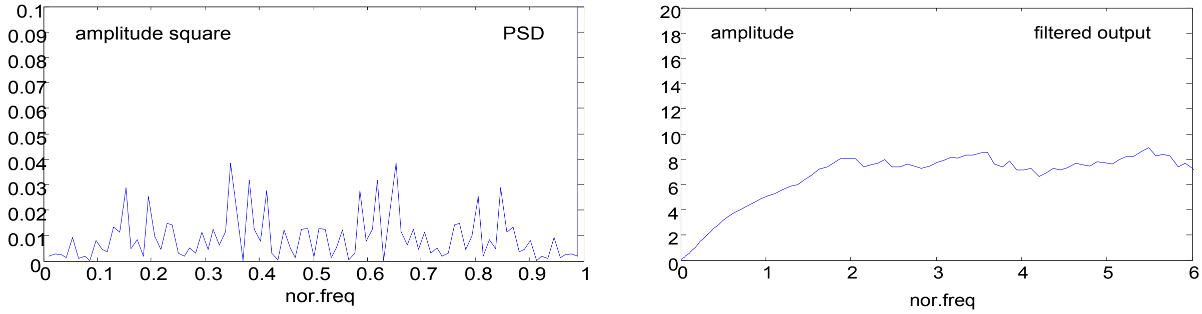
Without Filter With Filter

Figure 5. DFT Power Spectrum Plot For Cancer Cell of accession no. NM_012403.1 and length of EXON 1-705bp



Without Filter With Filter

Figure 6. DFT Power Spectrum Plot For Non-Cancer Cell of accession no. AF18660.1 and length of EXON (1210-1301) bp



Without Filter With Filter

Figure 7. DFT Power Spectrum Plot For Non-Cancer Cell of accession no. AF083883 and length of EXON (1210-1301) bp

❖ CONCLUSIONS

DSP nowadays plays an important role in DNA sequence analysis, CANCER diagnosis and gene expression analysis etc. Researchers are using DFT power spectrum plot to predict protein coding regions of a DNA sequence. Here we surveyed DFT power spectrum as a method to predict CANCER disease for various databases available in Gene bank. The result shows this method can be used as an easy tool to predict cancer disease. The filtered power spectrum plots yield high accuracy. The bar plots show prominent results for prediction of cancer. The authors have conducted some preliminary studies about the prediction of CANCER cells. Further efforts will be made to improve the accuracy of prediction by using other types of digital filters. And the next step in our research would be to generalize this analysis for a number of other oncogene databases.

❖ REFERENCES

[1.] Anastassiou, D: DSP in genomics: Processing and frequency domain analysis of character strings. IEEE, 0-7803-7041-2001.
[2.] Akhtar, Ambikairajah M., E. & Epps, J: Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction. IEEE Journal of selected topics in signal processing, Vol.2, No.3, June 2008.
[3.] Chen Jie and Wong Stephen T.C.: Nanotechnology for Genomic signal Processing in cancer Research. IEEE Signal Processing Magazine, January, 2007.
[4.] Deergha.Rao K and Swamy M.N.S: Analysis of Genomics and Proteomics using DSP techniques. IEEE Transactions on circuits and Systems, Vol 55, No. 1. Pp 370-378, February 2008.
[5.] Dougherty Edward R. and Dutta Aniruddha: Genomic Signal Processing: Diagnosis and Therapy. IEEE Signal Processing Magazine , January, 2005
[6.] Ficket, J.W., Tung, and C.S: Recognition of protein coding regions in DNA sequences. Nucleic Acids Research, Vol.10, No.17, pp.5303-5318, July 1982.
[7.] Ficket, J.W. & Tung, C.S: Assessment of protein coding measures. Nucleic Acid Research, , Vol.20, N0.24, 6441-6450,1992.
[8.] Herzel, H. & Grobe, B: Measuring correlations in symbol sequences. Phys. A, Vol. 216, pp.518-542, 1995.
[9.] Li, W. & Kaneko, K: Long range correlation and partial 1/f spectrum in a non-coding DNA sequence. Europhys. Lett., Vol.17, No.7, pp.655-660, January 1992.
[10.] Peng Qiu, Wang Z.Jane, and K.J. Ray Liu: Genomic Processing of Cancer Classification and Prediction. IEEE Signal Processing Magazine, January, 2007.
[11.] M.Roy, S.Biswas and S.Barman(Mandal): Identification and analysis of coding and non-regions of a DNA sequence by Positional Frequency Distribution of Nucleotides (PFDN) algorithm. International Conference on Computers and Devices for Communication (CODEC-09), December 14-15, 2009.
[12.] Tuqan. J. and Rushdi. A:A DSP based approach for finding the codon bias in DNA sequences. IEEE journal on signal processing, vol.2.No. 3, June, 2008.
[13.] Vierra, M.S: Statistics of DNA sequence: A low frequency analysis. Physical Review, Vol.60, No.5, Nov.1995.
[14.] Voss, R.F: Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phy. Rev. Lett., Vol.68, No.25, pp.3805-3808, 1992.
[15.] Hayes. M.H:Statistical digital signal processing and modelling. John Wiley & Sons, Inc., New York, USA,1996
[16.] National Centre for Biotechnology Information (NCBI). [Online]. Available: http://www.ncbi.nlm.nih.gov/.