International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-10, August 2019

A DSP Technique for Prediction of Cancer Cell

Malaya Kumar Hota

Abstract*: At present cancer is an alarmed disease. According to medical research, central cause of cancer is due to the genetic abnormality. Most cancers are generated due to permanent change in the deoxyribonucleic acid (DNA). For the past two decades, genomic signal processing (GSP) is a vital area of research. It has engrossed the consideration of digital signal processing (DSP) researchers for the massive amount of data accessible in the public data base. By finding out the DNA sequence for cancer cells & normal cells of human beings & applying some digital signal processing (DSP) approaches on both, difference between them can be found. Previously, discrete Fourier transform (DFT) power spectrum was used to predict cancer cells of a DNA sequence. In this paper, discrete cosine transform (DCT) and discrete sine transform (DST) approaches are presented as an alternative to analyze the spectral characteristics of cancer cells and normal cells. Further, post-processing is done using digital IIR low pass filter to improve the discrepancy between cancer and normal cells. The proposed method is tested for a number of data sets available in Gene Bank.

***Keywords:** Digital Signal Processing (DSP), Digital IIR low pass filter, Discrete Cosine Transform (DCT), Discrete Sine Transform (DST).*

I. INTRODUCTION

Currently cancer is a frightened disease that plays a primary role triggering death entirely over the sphere. It is caused due to abnormal proliferation of cells in a particular organ of the body with the potential to spread into other vital organs and infect to any other part of the body. Now why the cells proliferate abnormally is a question that must be answered. The presence of cancer-causing genes in the cells is a prerequisite for developing cancer. These cancer-causing genes known as oncogenes are either present in a person originally or they may be formed due to mutation of normal genes. There are many factors which may convert normal genes to mutated genes. A good example is the continuous use of tobacco over a long period, caused damage to the mucous membrane of the area of contact due to chronic irritation and lead to gene mutation ultimately leading to cause of cancer. We are living in a society where industrial pollution, excess use of pesticides in foods and exhaust from automobiles are more conducive for the development of cancer.

Cancer is mainly initiated due to the irregularities in the hereditary material of the converted cell. These irregularities may haphazardly occur due to mistakes in deoxyribonucleic acid (DNA) replication [1]. Sometimes these irregularities are inborn and accordingly existing in all cells from natal [1]. Most cancers are due to random mutations when cells are going through cell division. Sometimes cancers occur due to exposure to radioactivity or chemicals.

In the recent era, the study of genomics using digital signal processing (DSP) techniques have gained countless admiration. Since irregularity of the DNA and exon regions are allied to cancer. Therefore the focus of this work is to study the spectral characteristic of coding regions using DSP techniques.

A. Brief overview of DNA

A DNA is a twofold helix structure. It comprising of dual paired strands of sugar-phosphate backbone with nucleotide attached to it. The series of nucleotide *A*, *T*, *C* and *G* (respectively, adenine, thymine, cytosine, and guanine) provide all genetic information. One strand of DNA sequence can be divided into genes and inter-genic spaces in which genes are responsible for protein synthesis. Living organisms can be categorized into two types: *prokaryotes* and *eukaryotes*. Most prokaryotes are unicellular and cell nucleus is absent. While eukaryotes are multicellular with membrane-bound nuclei. In the complex organism (eukaryotes), the genes can further be split into exons and introns. Exons of a DNA sequence are primarily information bearing portion because only the exons take part in protein synthesis [2]. A mutation is a undying alteration in the DNA which can rise impulsively. Further, the mutation can arise in response to radiation from UV light or exposure to certain chemicals. Due to mutation sometimes huge segments are altered. This can lead to cancer. Previously, power spectrum plot using DFT was used to predict cancer cells of a DNA sequence. In this work, DCT and DST approaches are presented as an alternative to studying the spectral characteristics of cancer cells and normal cells. Further, digital IIR low pass filter post-processing is done to analyze the difference between cancer and normal cells.

B. Causes of Cancer

Cancer is an abnormal growth of cells. Cancer cells rapidly repeat and are often shaped differently from healthy cells. They can spread to several areas of the body. Oncology is the study of cancer and tumors. When a tumor is malignant, the term “cancer” is used. It has the potential to cause severe harm, including death. When a modification or mutation in a gene is occurred in germ cells, it is referred to as a “germline mutation”. This modification is inherited and are involved in a small percentage of causing cancer. On other hands, a mutation that takes place by chance over time in cells of the body is said to be “acquired” which are not passed down to our children. The prime cause of cancer is acquired mutation.

Manuscript published on 30 August 2019.
*****Correspondence Author(s)
Malaya Kumar Hota, Professor, Department of Communication Engineering, School of Electronics Engineering, Vellore Institute of Technology, Vellore 632014, Tamil Nadu, India.

The mutation of different types of genes leads to the development of cancer.

C. Coding and Noncoding Regions

In recent times scholars from diverse areas have focused in the arena of DNA sequence analysis. The main reason is the massive information content concealed in it [3]. The actual information bearing part of a DNA sequence is the exon regions. Therefore, the study of this part is of key importance. Today DSP plays a vital role in this effort because the DSP technique can identify hidden periodicities and features which cannot be revealed easily by conventional statistical methods. It has been established that the protein coding (exon) regions of DNA molecules exhibit a period-3 property which is also called three base periodicity [4]. This property is absent in the intron regions and inter-genic spaces. Therefore, this property is used as the discrimination of coding and non-coding regions. The basis of this property is consequent from the triplet nature of codon. However, this fact only is insufficient to describe why the coding regions exclusively have the period-3 component. Some researchers have attributed this property to the non-uniform usage of codons. This is commonly referred to as the *degeneracy of the genetic code*.

D. Prediction of Cancer using DSP technique

Previously discrete Fourier transform (DFT) power spectrum was used for prediction of cancer cell [5] by representing the signals in the frequency domain. In this work, instead of DFT, discrete cosine transform (DCT) & discrete sine Transform (DST) are used as an alternative.
DCT operates on real data with even symmetry. It uses only (real-valued) cosine functions. The DST uses only (real-valued) sine functions which is equivalent to the imaginary parts of a DFT.

In DNA sequences, the nucleotide bases are represented by a string of character. The first step in DSP techniques is called mapping which is used to convert the character strings into numerical sequences [6]. Numerous mapping techniques are used by scholars to convert the string of character into numerical sequences. It has been shown in [4] that among various mapping techniques the prediction accuracy is maximum in Voss, Z-curve, and tetrahedron mapping techniques. Almost similar result is obtained in all three techniques. In z-curve method, the DNA sequence is converted into three-dimensional representations based on the symmetry of the regular tetrahedrons. The Z-curve contains all the information that the corresponding DNA sequence carries. In the tetrahedron mapping technique, the number of indicator sequences also reduces from four to three in a manner symmetric to all four components. Due to simplicity, Voss mapping technique is preferred in this work. This is perhaps the initial mapping technique. In this mapping, four indicator sequences are generated based on the presence or absence of that particular nucleotide. That is, the DNA sequence is mapped into four indicator sequences $x_i[n]$, $\forall\, i \in F = \{A, C, G, T\}$. The binary indicator sequence $x_i[n]$ of length N takes the value of 1 or 0 depending on the presence or absence of base i at position n. If the DNA sequence is "*A T A G T C A G C T*", then $x_A[n] = 1\ 0\ 1\ 0\ 0\ 0\ 1\ 0\ 0\ 0$, $x_T[n] = 0\ 1\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 1$, $x_C[n] = 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1\ 0$ and $x_G[n] = 0\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ 0\ 0$.

The DFT of the indicator sequence is

$$X_{DFT}[k] = \sum x_{(A,T,C,G)}[n]\, e^{-\frac{j2\pi kn}{N}} \qquad (1)$$

where k=0, 1, 2……N-1 and n=0, 1, 2……N-1.

The DCT of the indicator sequence is

$$X_{DCT}[k] = \omega(k) \sum_{n=1}^{N} x_{(A,T,C,G)}[n] \left(\cos \frac{\pi(2n-1)(k-1)}{2N}\right) \qquad (2)$$

where $k = 1,2,3 \ldots\ldots\ldots\ldots\ldots\ldots N$ and

$$\omega(k) = \begin{cases} \frac{1}{\sqrt{N}} & k = 1 \\ \sqrt{\frac{2}{N}} & 2 \le k \le N \end{cases}$$

The DST of the indicator sequence is

$$X_{DST}[k] = \sum_{n=1}^{N} x_{(A,T,C,G)}[n] \sin\left(\pi \frac{kn}{N+1}\right) \qquad (3)$$

where $k = 1,2,3 \ldots\ldots\ldots\ldots\ldots\ldots N$

Then the Power Spectral of the DFT sequence is given by

$$P_s[k] = \sum |X_{DFT}[k]|^2 \qquad (4)$$

The Power Spectral of the DCT sequence is given by

$$P_s[k] = \sum |X_{DCT}[k]|^2 \qquad (5)$$

The Power Spectral of the DST sequence is given by

$$P_s[k] = \sum |X_{DST}[k]|^2 \qquad (6)$$

The plot of power spectral of the exon region is explored as a measure of the cancer cell or normal cell. In this paper, a digital IIR low pass filter with an elliptic approximation is used to suppress the noise from the power spectrum. Due to this post processing using IIR low pass filter, a better prediction of cancer cells is possible from non-cancer cells. In the elliptic filter both passband and stopband are equiripple. In this work elliptic filter is used, because equiripple response requires a reduced order for the identical set of ripple sizes and transition bandwidth. In this work, a low pass digital elliptic filter of order 10 with passband edge frequency 0.3607, peak-to-peak ripple 0.5 dB and minimal stopband attenuation 40 dB is selected. Empirically it is found that these parameters are most suitable for improved prediction of the cancer cell.

II. PROPOSED METHOD FOR PREDICTION OF CANCER

Previously DFT power spectrum was used for prediction of cancer cell [5]. In this paper, instead of DFT, filtered power spectrum plot using DCT and DST is used for better prediction of the cancer cell. Fig. 1 shows the schematic data flow diagram of the proposed method.
The method has following steps:

Step 1: numerical mapping of the DNA sequence into indicator sequences based on Voss mapping technique,

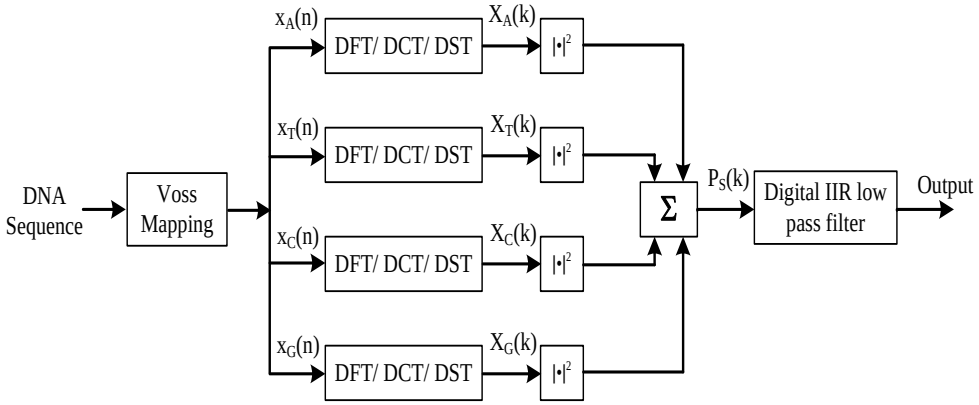


Fig. 1. Schematic data flow diagram of the proposed method for cancer cell prediction

Step 2: the application of DFT/ DCT/ DST to each indicator sequence,

Step 3: obtain the power spectral content of the DFT/ DCT/ DST sequence,

Step 4: filter the power spectrum of the sequence using a digital IIR low-pass filter.

III. RESULT AND DISCUSSION

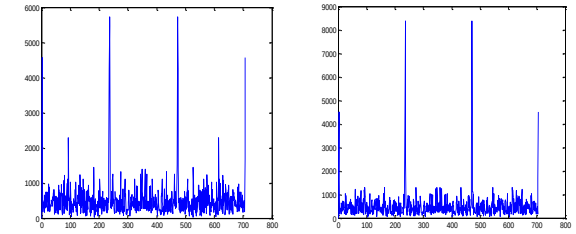
Although, for prediction of cancer cell, previously DFT power spectrum was used. But in this paper, DCT power spectrum and DST power spectrum are used as an alternative to predict the discrepancy among cancer cell and non-cancer cell. For testing the prediction accuracy several data bases are considered for our test in this work.

Some of these databases with accession numbers are shown in Table I [5].

Table- I: Accession numbers of cancer cells and normal cells [5]

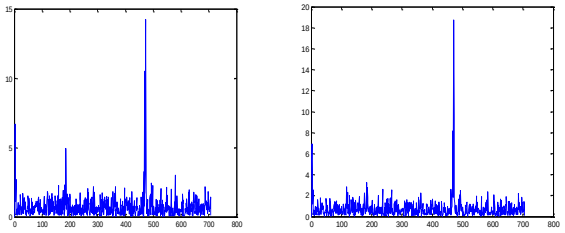
Sl. No.	Type of Cells	Accession No.
1	Cancer cells	NM_012403.1
2		AF348515.1
3		NM_016346.2
4		NM_005732.3
5		AF348525.1
6		AF008216.1
7	Normal cells	AF083883
8		AF186607.1
9		AF186613.1
10		AF007546

The DFT spectrum of the exon region is shown in Fig.2 for cancer cell. Similarly DCT spectrum and DST spectrum of the exon region are shown in Fig.3 and Fig.4 respectively for cancer cell. Further, for non-cancer cell the DFT spectrum, DCT spectrum and DST spectrum are plotted in Fig. 5, Fig.6 and Fig.7 respectively.



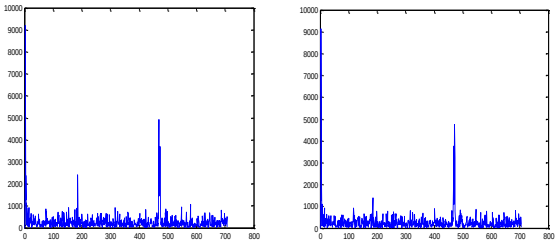
(a) **(b)**

Fig. 2. DFT power spectrum for cancer cell (a) AF008216 (exon 4453-5157 bp) and (b) NM_012403 (exon 1-705 bp)



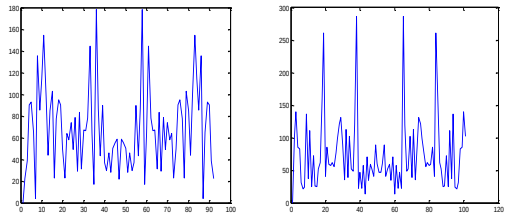
(a) **(b)**

Fig. 3. DCT power spectrum plot for cancer cell (a) AF008216 (exon 4453-5157 bp) and (b) NM_012403 (exon 1-705 bp)



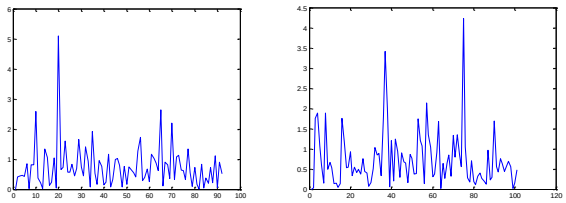
(a) **(b)**

Fig. 4. DST power spectrum plot for cancer cell (a) AF008216 (exon 4453-5157 bp) and (b) NM_012403 (exon 1-705 bp)



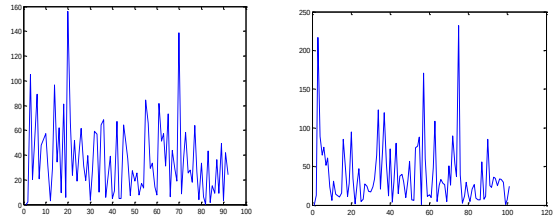
(a) **(b)**

Fig. 5. DFT power spectrum plot for non-cancer cell (a) AF186607.1 (exon 1210-1301 bp) and (b) AF083883 (exon 1210-1301 bp)



(a) **(b)**

Fig. 6. DCT power spectrum plot for non-cancer cell (a) AF186607.1 (exon 1210-1301 bp) and (b) AF083883 (exon 1210-1301 bp)

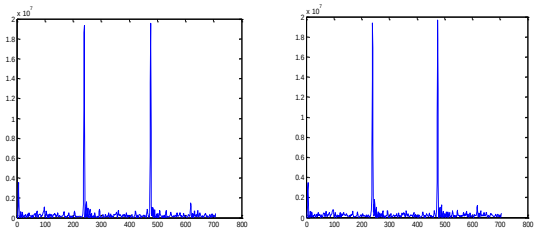


(a) **(b)**

Fig. 7. DST power spectrum plot for non-cancer cell (a) AF186607.1 (exon 1210-1301 bp) and (b) AF083883 (exon 1210-1301 bp)

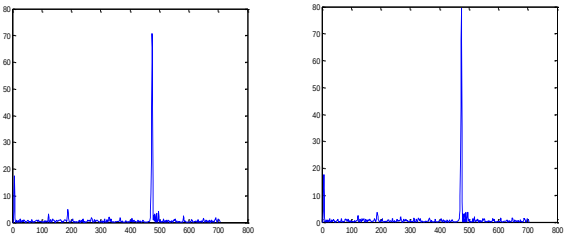
From the DFT, DCT, DST power spectrum plot shown in Fig. 2 to Fig.7 of the exon region, it can be understood that there is a clear discrepancy between a cancer cell and normal cell. Spikes are generated in the spectrum of cancer cells, which is absent in normal cells. Spikes are exactly at one-third of the length of exon.

In this work, DCT and DST approaches are presented as an alternative to studying the spectral characteristics of cancer cells and normal cells. Further, post-processing is done using digital IIR low pass filter to analyze the difference between cancer and normal cells. The post processing using IIR low pass filter is done in this work to suppress the high frequency noise from the power spectrum. The IIR filter output of exon region is shown in Fig. 8 to Fig. 10 for cancer cell. Similarly, the IIR filter output of exon region is shown in Fig. 11 to Fig. 13 for normal cell. It is depicted from Fig. 8 to Fig. 13 that the IIR filter output gives a strong dissimilarity between cancer cell & normal cell. By plotting the DFT, DCT and DST power spectrum post processing with IIR low pass filter, we observed some difference between both the cells. It is found that there is a strong discrepancy between the cancer cell and normal cell.



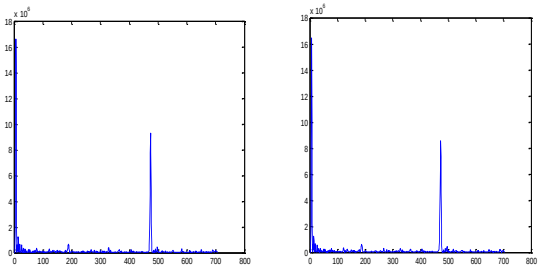
(a) **(b)**

Fig. 8. DFT power spectrum post processing with IIR low pass filter for cancer cell (a) AF008216 (exon 4453-5157 bp) and (b) NM_012403 (exon 1-705 bp)



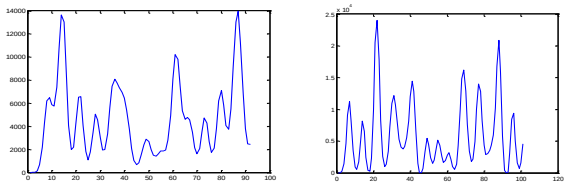
(a) **(b)**

Fig. 9. DCT power spectrum post processing with IIR low pass filter for cancer cell (a) AF008216 (exon 4453-5157 bp) and (b) NM_012403 (exon 1-705 bp)



(a) **(b)**

Fig. 10. DST power spectrum post processing with IIR low pass filter for cancer cell (a) AF008216 (exon 4453-5157 bp) and (b) NM_012403 (exon 1-705 bp)

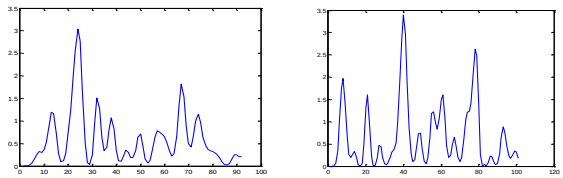


(a) **(b)**

Fig. 11. DFT power spectrum post processing with IIR low pass filter for non-cancer cell (a) AF186607.1 (exon 1210-1301 bp) and (b) AF083883 (exon 1210-1301 bp)

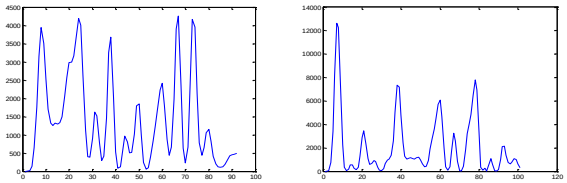
Retrieval Number: F3409048619/19©BEIESP
DOI: 10.35940/ijitee.F34092435.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)



(a) (b)

Fig. 12. DCT power spectrum post processing with IIR low pass filter for non-cancer cell (a) AF186607.1 (exon 1210-1301 bp) and (b) AF083883 (exon 1210-1301 bp)



(a) (b)

Fig. 13. DST power spectrum post processing with IIR low pass filter for non-cancer cell (a) AF186607.1 (exon 1210-1301 bp) and (b) AF083883 (exon 1210-1301 bp).

In Fig. 14, a comparative plot between DCT & DFT power spectrum is shown. In Fig. 15 a comparative plot between DCT & DFT power spectrum post processing with IIR low pass filter is shown. From Fig. 14 and 15 it is observed that the spikes are more prominent in the power spectrum plot of DCT compared to DFT. The discrimination between the spikes and other regions is further increased when they are post-processing using IIR low pass filter.

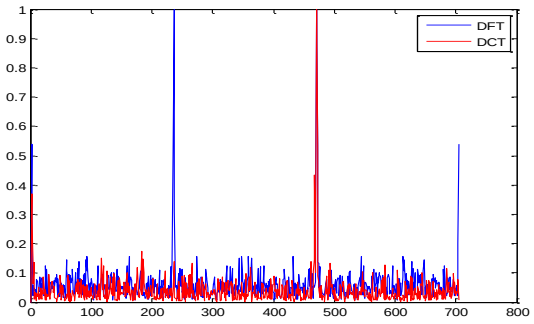


Fig. 14. DFT &DCT power spectrum plot for NM_012403

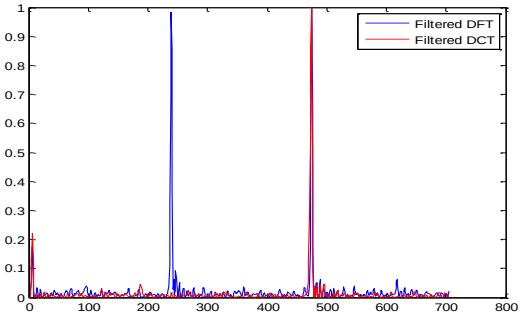


Fig. 15. DFT & DCT power spectrum post processing with IIR low pass filter plot for NM_012403

IV. CONCLUSION

These days DSP plays a significant role in genomic sequence study and prediction of cancer cell. Previously, DFT power spectrum plot was used to predict cancer cells of a DNA sequence [5]. In this paper, DCT and DST power spectrum is proposed as alternative methods to predict cancer disease for several databases available in Gene-bank. Further, IIR low pass filter with elliptic approximation is used for suppression of noise and improvement of discrepancy. The result revels that the proposed method can be used as an alternative to predict cancer disease. It is found that the filtered power spectrum plots yield strong discrepancy between the cancer cell and normal cell. Furthermore, the filtered power spectrum plot using DCT technique is more accurate than DFT & DST.

REFERENCES

1. Q. Peng, Z.J. Wang, and K.J.R. Liu, "Genomic processing for cancer classification and prediction — a broad review of the recent advances in model-based genomic and proteomic signal processing for cancer detection," *IEEE Signal Processing Magazine*, Vol. 24, No. 1, 2007, pp. 100-110.
2. M. Akhtar, E. Ambikairajah, and J. Epps, "Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction," *IEEE Journal of selected topics in signal processing*, Vol. 2, No.3, 2008, pp. 310-321.
3. J. Tuqan, and A. Rushdi, "A DSP based approach for finding the codon bias in DNA sequences," *IEEE Journal of Selected Topics in signal processing*, Vol. 2, No. 3, 2008, pp. 343-356.
4. M.K. Hota, and V.K. Srivastav, "Identification of protein coding regions using anti-notch filters," *Digital signal processing, Elsevier*, Vol. 22, No. 6, 2012, pp. 869-877.
5. S. Barman, M. Roy, S. Biswas, and S. Saha, "Prediction of Cancer Cell Using Digital Signal Processing," *International Journal of Engineering*, Vol. 9, No. 3, 2011, pp. 91-95.
6. T. Meng, A. T. Soliman, and M.-L. Shyu et al., "Wavelet analysis in current cancer genome research: a survey," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 10, No. 6, 2013, pp. 1442–1459.

AUTHORS PROFILE

Malaya Kumar Hota is the Professor in the department of Communication Engineering, School of Electronics Engineering at Vellore Institute of Technology (VIT), Vellore, Tamilnadu, India. He was previously a Professor and Principal at SIET, Dhenkanal, Odisha. He has more than sixteen years of teaching and research experience. He received his M.Tech. in Electronics Engineering from Visvesvaraya NIT, Nagpur, India, in 2002 and Ph.D. in Electronics and Communication Engineering from Motilal Nehru NIT, Allahabad, India, in 2011. He has authored or co-authored about twenty-two publications. He received one MODROBS grant from AICTE for Modernization of Digital Signal Processing Lab. His biography has been included in Marquis Who's Who in Science and Engineering, and also in Marquis Who's Who in the World. His main research interest is in digital signal processing, genomic signal processing, biomedical signal processing and non-stationary signal processing.