POWER SPECTRUM SEQUENCE ANALYSIS OF RHEUMATIC ARTHRITIS (RA DISEASE USING DSP TECHNIQUE

K.B.Ramesh[1], Prabhu Shankar.K.S[2], B.P.Mallikarjunaswamy[3], E.T.Puttaiah[4]

[1]*Associate Professor, Dept. of Instrumentation Technology, R.V.College of Engineering, Bangalore, India*
[2]*Biomedical Signal Processing and Instrumentation, I.T Dept., R.V.College of Engineering, Bangalore, India*
[3]*Professor, Department of Computer Science and Engineering, SSIT, Tumkur, Karnataka, India*
[4]*Professor, Dept. of Environmental science, Vice-chancellor, Gulbarga University, Karnataka, India*
ramesh_k_b@hotmail.com, prabhushankar.k.s@gmail.com, drbpswamy@rediffmail.com, profetputtaiah@rediffmail.com

Abstract

Digital Signal Processing (DSP) applications in bioinformatics have received great attention in recent years, where new effective methods for genomic sequence analysis, such as the detection of coding regions, have been developed. Rheumatic Arthritis (RA) is a chronic systemic inflammatory disease involving primarily the peripheral synovial joints. In this work, the software module has been implemented using MATLAB 2009a which supports bioinformatics toolbox. The DSP techniques such as Fast Fourier Transform (FFT) and Hamming window are incorporated in the algorithm. Quantitative analysis is performed using mean amplitude and mean normalized frequency parameters computed from the generated power spectrum. The algorithm is tested for different normal and abnormal DNA sequences available in National center of Biotechnology Information (NCBI) database.

Keywords: *Digital signal processing (DSP), Fast Fourier Transform (FFT), Rheumatic Arthritis (RA).*

----------***----------

1. INTRODUCTION

Bioinformatics is an interdisciplinary field that develops and improves upon methods for storing, retrieving, organizing and analyzing biological data. A major activity in bioinformatics is to develop software tools to generate useful biological knowledge. Bioinformatics tools aid in the comparison of genetic and genomic data and more generally in the understanding of evolutionary aspects of molecular biology. The process of predicting genes in the field of bioinformatics has been traditionally done and is often seen as being long and expensive.

Genomic Signal Processing is a fundamental discipline that brings to genomics the structural model-based analysis and synthesis that form the basis of mathematically rigorous engineering. Signals generated by the genome must be processed to characterize their regulatory effects and their relationship to changes at both the genotypic and phenotypic levels. Application is generally directed towards tissue classification and the discovery of signaling pathways. Because transcriptional control is accomplished by a complex method that interprets a variety of inputs, the developments of analytical tools that detect multivariate influences on decision making present in complex genetic networks are essential. To carry out such an analysis, one needs appropriate analytical methodologies. Authors suggested several DSP techniques and methods to predict the protein coding regions of gene but still there is requirement for developing effective modules for the identification of protein coding regions in the DNA sequences pertaining to several diseases and for RA in particular. Soft computing methods neural network, genetic algorithms, fuzzy logic for predicting exon regions generate superior results but in spite of efficacy of these methods, their system's implementation is complex and thus cannot be used widely laboratories. The existing DSP techniques filtering techniques, DFT technique, modified Gabor wavelet method and a method based on average magnitude difference function(AMDF) suffers from one of the problems like inaccurate in finding the protein coding regions due to addition of noise, requires more processing time, and complexity. And also string matching methods like dynamic programming and heuristic techniques like BLAST was used to find exon regions but was not able obtain the satisfactory results.

Rheumatic Arthritis (RA) is a chronic, progressive and disabling auto-immune disease. Around 2-4% of world's population and 0.6% of India are suffering from RA. It is a painful condition and can cause severe disability and ultimately affects a person's ability to carry out everyday tasks. The diagnosis of early RA is very difficult. It is the goal of every rheumatologist to try to prevent joint damage, the earlier and

more aggressively someone with RA is treated, the better the long term prognosis.

The disease is progressive and results in pain, stiffness, and swelling of joints, which shows deformity (deviation of joints) and ankylosis (stiffness at joints) in the late stages of the disease. Recurring inflammation of the affected joints (i.e., arthritis) leads to a degradation of cartilage and to erosive destructions (erosions) of the bone as shown in figure 1 compared to normal joint.

Signal processing techniques have been widely used in the last decade in gene prediction and the genomic signal processing

(GSP) field received a consistent effort from researchers. The objective of the proposed work is to develop a genomic software module using MATLAB to detect and predict the presence of RA. The generated results will assist doctors for the characterization of disease, medical practioners and research community for better understanding of the disease development.

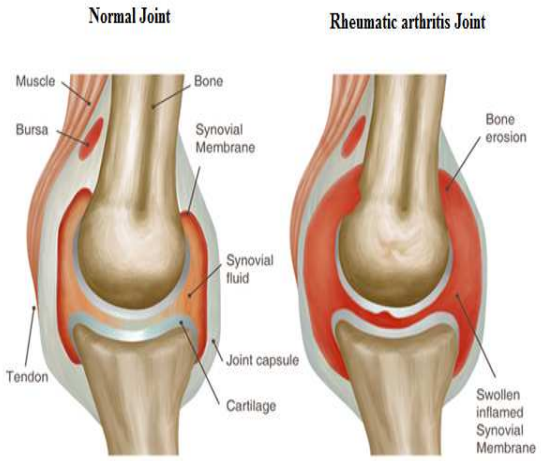

Figure 1: Normal and Arthritic joint

This developed software module can be used by the doctors for the characterization of the disease and to identify the cause for disease so that proper diagnosis can be implied to the patients. It is useful for the Pharmacists to discover new medicines for rheumatic arthritis disease, for research communities to understand the factors and genes responsible for patients with rheumatic arthritis.

Deoxyribo Nucleic Acid (DNA) is made up of linear chains of subunits called nucleotides. Each nucleotide consists of three parts; a nitrogenous base, a five-carbon-atom sugar and a phosphate group. The four possible nucleotide bases are adenine (A), cytosine (C), guanine (G), and thymine (T). However, in a number of viruses, where ribonucleic acid (RNA) is used as building block of their genome, thymine is replaced by uracil (U). In DNA, individual nucleotides are connected to each other through sugar-phosphate bonds, forming a long one dimensional chain with two distinct ends, the 5' end (upstream), and the 3' end (downstream).

Therefore, this DNA chain or strand can symbolically be represented by a character string consisting of four alphabet letters A, C, G, and T. The DNA double-helix model (Figure 3.4), which was originally proposed by J. D. Watson and F. C. H. Crick [14], holds the following interesting features [15]:

- Two DNA strands coiled around a central axis form a right-handed double helix i.e., their turns are clockwise when looking down the helical axis.
- The two strands are antiparallel i.e., each strand has specific orientation and they run in opposite directions.
- The nitrogen bases of opposite strands are hydrogen bonded.
- The purine of one strand is paired with pyrimidine of the other strand i.e., A is paired to T and vice versa, and G is linked to C and vice versa.
- The purine-pyrimidine pairs are being complement to each other, thus, two strands of a single DNA molecule are complementary to one another. Thus, if sequence 5' – CATTGCCAGT – 3' occurs on one strand, the other strand must have the sequence 5' – ACTGGCAATG – 3'

i.e.,

→ Strand one: 5' – C A T T G C C A G T – 3'

→ Strand two: 3' – G T A A C G G T C A – 5'

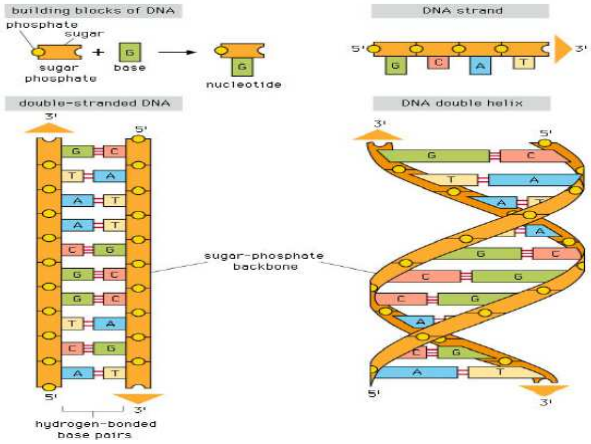

Figure 2: DNA and its building blocks [16]

A DNA sequence can be divided into genes and intragenic spaces. The genes are responsible for protein synthesis. A gene can be divided into two sub regions called the exons and introns as shown in Figure 3.6.

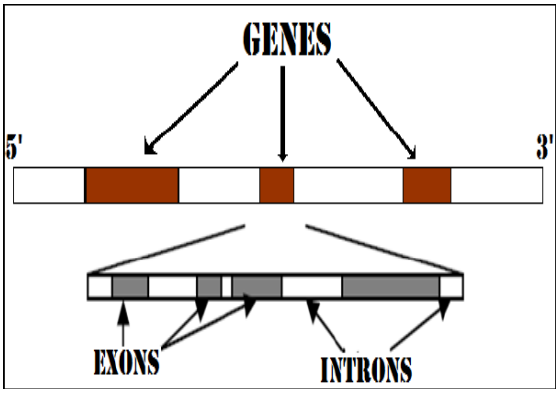

Figure 3: Various regions in a Gene.

Only the exons are involved in protein coding. The bases in the exon region can be divided into groups of three adjacent bases. Each triplet is called a codon. Scanning the gene from left to right, a codon sequence can be defined by concatenation of the codons in all the exons. Each codon instructs the cell machinery to synthesize an amino acid. The codon sequence therefore uniquely identifies an amino acid sequence which defines a protein.

2. METHODOLOGY

FFT–DSP method is the simple method used to perform comparative analysis of both the gene sequences of a person with RA and for a non RA. The FFT technique is a DSP based approach which is implemented using MATLAB. The tool accepts these inputs from the user and performs the DSP operations on it, and provides the user with the output power spectrum.

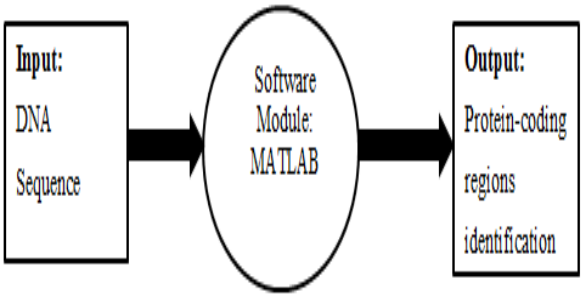

Figure 4: Basic Block Diagram of the System

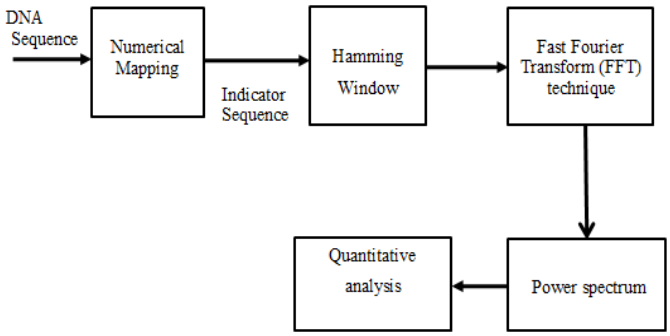

Figure 5: Functional Block diagram

The block diagram below depicts the steps involved in the implementation of the project. The DNA sequence is downloaded from the standard database NCBI of both normal and RA sequences. The process is broken into four main components.

The first component in the process is to convert symbolic DNA sequences into numerical sequences since DSP tools can process only numerical entities. The second step involves in the processing of sequences based on a window, and the window shape and length are the important parameters. The third stage, the analysis tool, is the most important component in the process. In this step, the period-3 component is extracted from DNA sequences to differentiate between coding and non-coding regions. Finally, the last step is the classification of exons and introns based on the threshold value.

2.1 Complex Indicator Sequence

The four binary indicator sequences can also be replaced by just one indicator sequence which is called as 'Complex indicator Sequence', xcis (n). For a position i in the DNA sequence x(n), the complex indicator sequence xcis(n) values are defined as:
If x (i) = A then xcis (i) = 1
x (i) = G then xcis(i) = -1
x (i) = T then xcis(i) = j
x (i) = C then xcis(i) = -j

In this mapping we have kept the purine (A and G) in the real axis and pyrimidine (T and C) in the imaginary axis. For example, if x (n) = ATGATCTGAA, then xcis (n) = [1 j -1 1 j -j j -1 1 1].

2.2 Hamming Window

When using FFT analysis to study the frequency spectrum of signals, there are limits on resolution between different frequencies, and on detectability of a small signal in the presence of large one. In effect, the process of measuring a signal for a finite time is equivalent to multiplying the signal by

a hamming function of unit amplitude. The hamming size of 351 is chosen for the analysis and it is optimized to minimize the maximum (nearest) side lobe, giving it a height of about one-fifth that of the Hann window, a raised cosine with simpler coefficients.

$$w(n) = \alpha - \beta \cos\left(\frac{2\pi n}{N-1}\right)$$

With $\alpha = 0.54$ $\beta = 1 - \alpha = 0.46$ $N = 351$

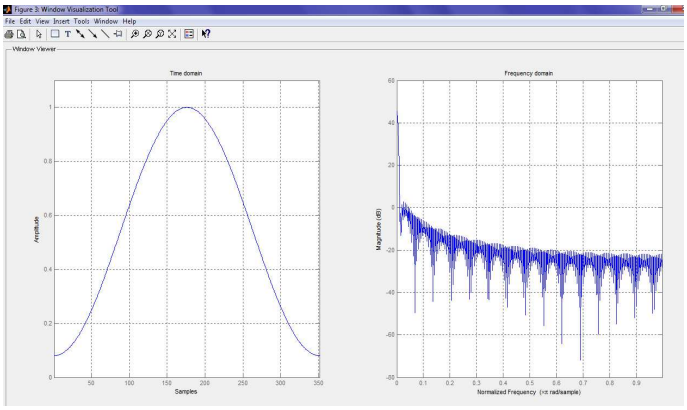

Figure 6: Impulse Response of Hamming Window.

2.3 Fast Fourier Transform

The FFT is a complex-valued linear transformation from the time domain to the frequency domain. An FFT computes the DFT and produces exactly the same result as evaluating the DFT definition directly; the only difference is that an FFT is much faster.

It involves taking FFT of a hamming window of a defined size in a DNA sequence, which is then slide across the whole length of the sequence [2].

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi k \frac{n}{N}} \qquad k = 0, \ldots, N-1.$$

The value of S (k) is large at k=N/3 when a coding region is present. We can calculate S (N/3) for short windows of data, and then slid the window by one or more bases and S (N/3) is recalculated. Thus we get a picture of how S (N/3) evolves along the length of the DNA sequence. Optimal window length of 300 is selected for better resolution and sharpness of the peak.

2.4 General Algorithm

Step 1: Retrieve the DNA sequence from the FASTA format using the Bio-Informatics tool box functions.
Step 2: Convert the DNA sequence into complex indicator sequence suitable for the spectral analysis.
Step 3: Create a hamming window of length 351, which is slide over whole length and s(n/3) periodicity is calculated.
Step 4: Compute the power spectral density by applying DSP technique, Fast Fourier transform to the sequences.
Step 5: Project the PSD values on the spectrum for the prediction of protein coding regions with some threshold co-efficient.
Step6: Calculate the Mean amplitude and Mean normalized Frequency of obtained each spectral plot.
Step7: Calculate the ratio of mean amplitude and mean normalized frequency for the classification of normal and RA sequence.

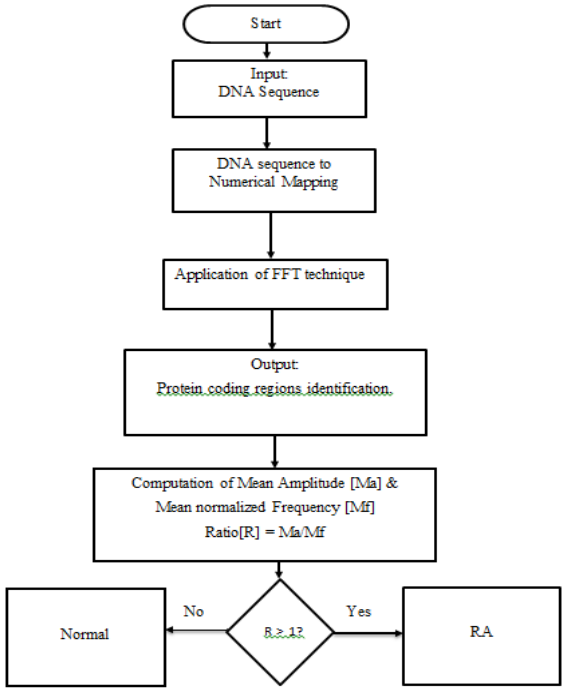

Figure 7: Flow chart of the proposed algorithm

3. RESULTS AND ANALYSIS

The software module for the detection of protein coding regions of a normal and abnormal (RA) sequences has been developed and tested for various cases. The snapshots and various test cases of gene sequence analysis for protein coding regions using DSP-FFT method for Complex indicator values are given in the following section

3.1 Results Obtained in Pre-Processing Process

Data extraction and Numerical mapping are the major steps in the process before applying the DSP technique to the sequence. DNA sequences of normal and RA status are taken from the standard database National center for Biotechnology Information (NCBI)[17] of Homo sapiens in the fasta format. The Downloaded sequence from NCBI is retrieved into the Matlab environment using the Bio informatics toolbox as shown in the below fig.4.1.

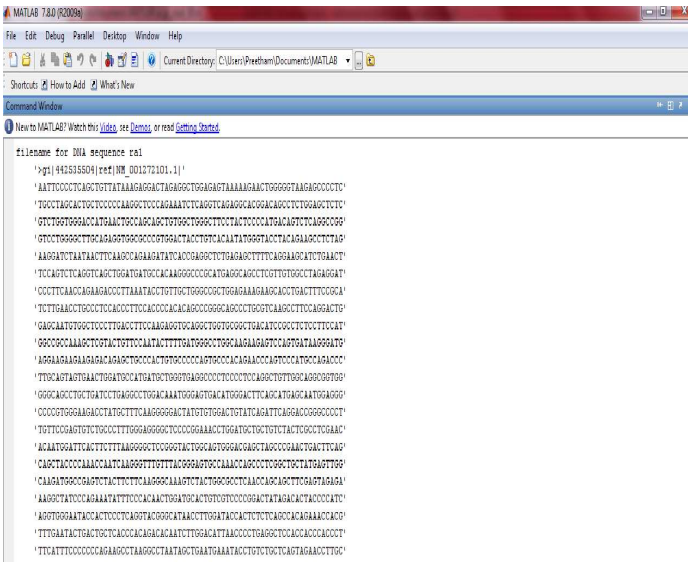

Figure8: Display of the DNA sequence retrieved from the fasta format into the Matlab using bioinformatics toolbox.

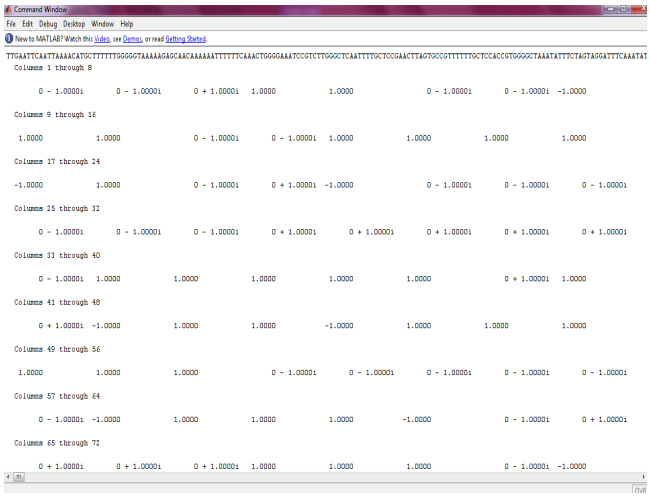

Figure 9: Display of the numerically mapped DNA sequence retrieved from the fasta format into the Matlab using complex indicator method.

3.2 ORF Finder

The ORF finder (Open Reading Frame) finder is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence already in the database.

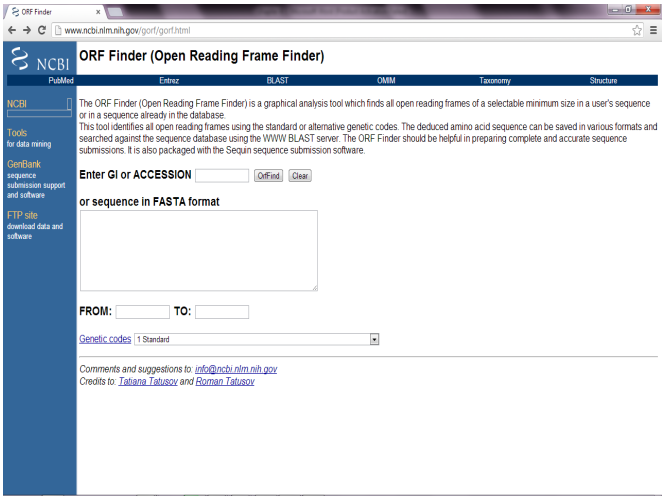

Figure 10: Snap shot of ORF Finder used to find the open reading frames of the DNA sequences.

This tool identifies all open reading frames using the standard or alternative genetic codes by just entering the reference no. in the tool shown in the fig. 4.3.The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the www.blastserver.com

3.3 Result analysis for Different Cases

Case 1
Reference No: NM_001199692.1
Length: 4943 bp
Generated Result:

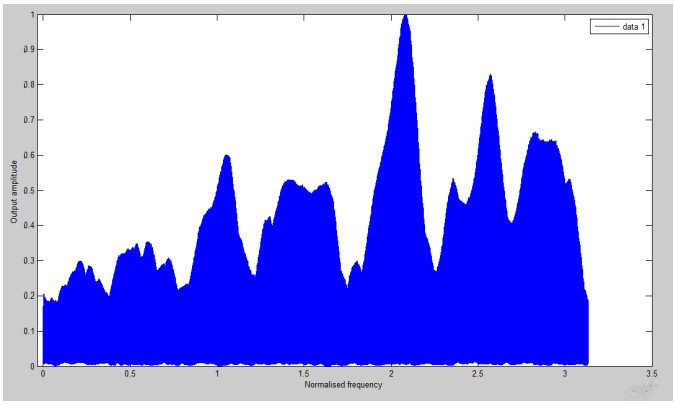

Figure 11: Output Spectrum obtained for case 1

X-axis: Position - length of the input sequence (Normalized Frequency)
Y-axis: Projection co-efficient (amplitude values) calculated as PSD (Power Spectral Density)
Mean Amplitude[Ma] = 2.6875
Mean Frequency[Mf] = 4.7
Ratio[R] = Ma/Mf = 0.519

Validation of Output for Case 1

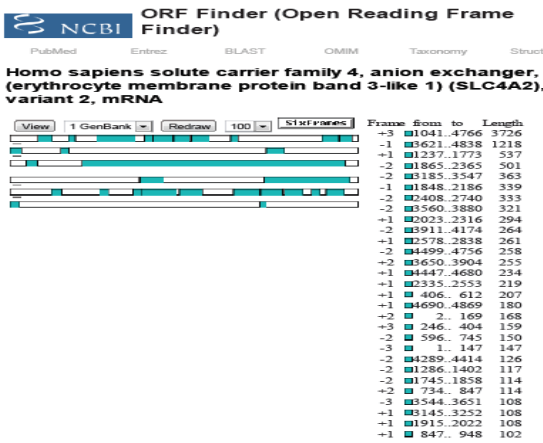

Figure 12: Validation of Output for case 1

Case 2
Reference No: NM_001184976.1
Length: 3780 bp
Generated Result:

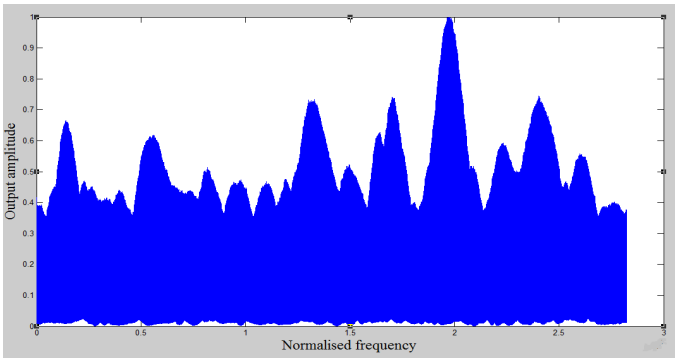

Figure 13: Output Spectrum obtained for case 2

X-axis: Position - length of the input sequence (Normalized Frequency)
Y-axis: Projection co-efficient (amplitude values) calculated as PSD (Power Spectral Density)
Mean Amplitude[Ma] = 3.65
Mean Frequency[Mf] = 3.67
Ratio[R] = Ma/Mf = 0.9

Validation of Output for Case 2

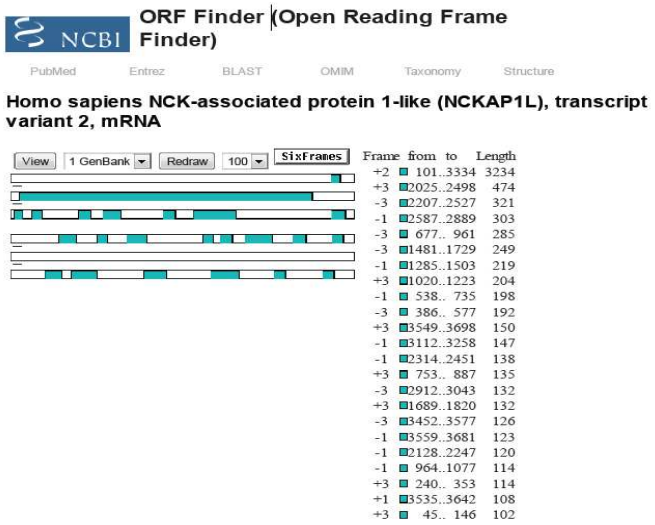

Figure 14: Validation of Output for case 2

Case 3
Reference No: XM_004317778.1
Length: 3035 bp
Generated Result:

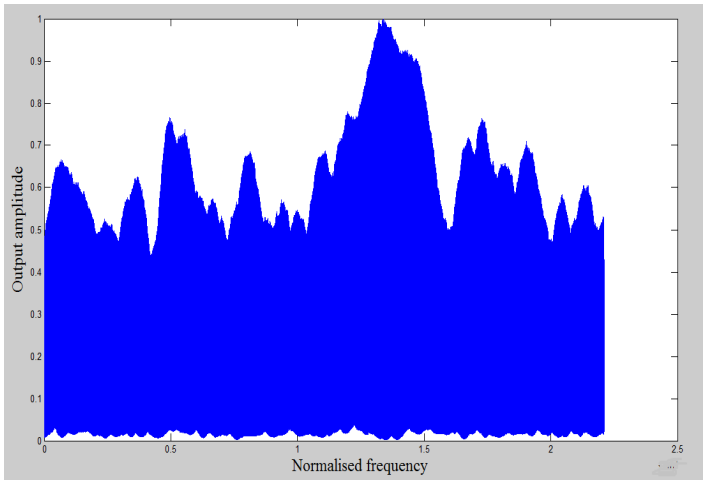

Figure 15: Output Spectrum obtained for case 3

X-axis: Position - length of the input sequence (Normalized Frequency)
Y-axis: Projection co-efficient (amplitude values) calculated as PSD (Power Spectral Density)
Mean Amplitude[Ma] = 3.575
Mean Frequency[Mf] = 3.978
Ratio[R] = Ma/Mf = 0.89

Validation of Output for Case 3

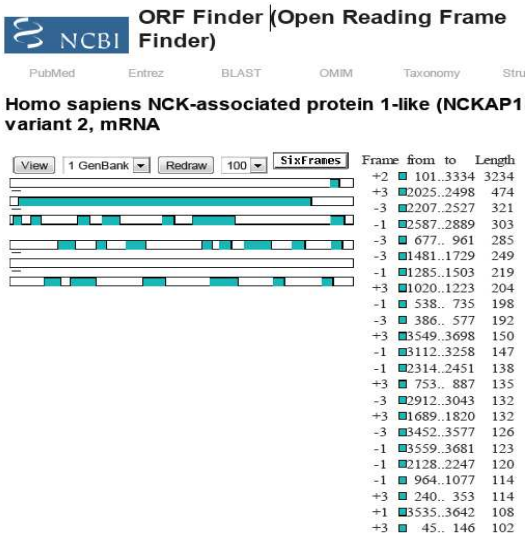

Figure 16: Validation of Output for case 3

The mean amplitude and mean normalized frequency and calculated ratio for all the downloaded sequences with their reference no.'s are tabulated in the below table 4.8

The computed ratio of mean amplitude to mean normalized frequency, less than 1.0 indicates the normal sequences and more than 1.0 for abnormal sequences is plotted in the bar plot as shown in the above figure 4.16

Table 1: Complete analysis of all 14 sequences with mean amplitude, normalized frequency and ratio[R]

Sequences	Mean amplitude [Ma]	Mean frequency [Mf]	Ratio = Ma/ Mf
Norm 1	2.6875	4.7	0.51
Norm 2	3.65	3.67	0.90
Norm 3	3.575	3.978	0.89
Norm 4	2.222	4.2812	0.51
Norm 5	2.5577	4.5220	0.56
Norm 6	2.6764	5.4444	0.49
Norm 7	2.7831	5.2352	0.53
RA1	3.942	3.5863	1.10
RA 2	3.5862	2.1101	1.69
RA 3	3.8997	3.0833	1.26
RA 4	5.8101	3.5862	1.62
RA 5	6.6698	4.0886	1.63
RA 6	4.439	3.083	1.43
RA 7	4.024	3.272	1.22

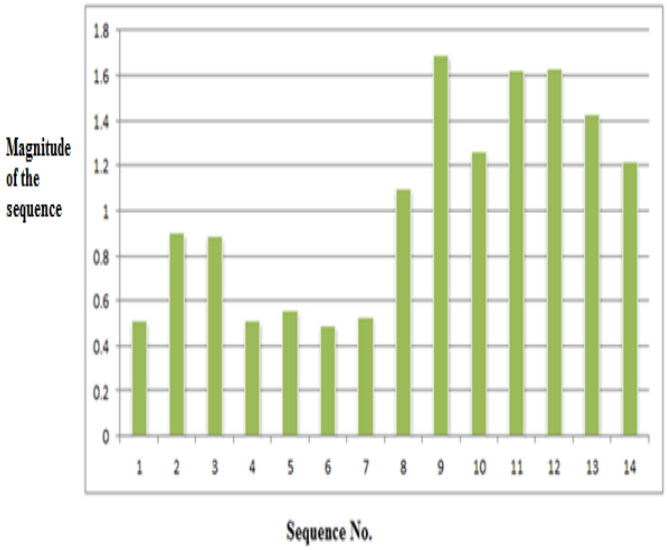

Figure 17: Plotted graph for all sequences.

CONCLUSION AND FUTURE SCOPE

Conclusion

- In the proposed work, software module has been developed using MATLAB which supports Bioinformatics tool box and DSP technique has been implemented and applied on the genomic data for the detection and prediction of protein coding regions and characterization of the disease.
- The proposed algorithm is efficient, requires less processing time, high accuracy for available genomic data and cost effective.
- The algorithm is tested for many different DNA sequences of both normal and abnormal (RA) available in National center of Biotechnology Information (NCBI) database.
- Ratio of mean amplitude to mean normalized frequency is computed so that, less than 1.0 indicates the normal sequences and more than 1.0 for abnormal sequences.

Future scope

- Further efforts can be made to improve the accuracy of the system.
- Efficiency of the developed module can be still improved by detecting the stage of disease.
- The proposed algorithm can be made as universal standard and also can be used to predict the other disease.

REFERENCES

[1] Vaidyanathan P.P and Yoon B.J "Digital filters for gene prediction applications," *IEEE Asilomar Conference. Signals Syst. Computers*, pp.306–310, Nov 2002.

[2] Vaidyanathan P.P and Yoon BJ "The role of signal-processing concepts in genomics and proteomics" *Journal of the Franklin Institute, special issue on Genomics*, vol. 341, no. 53, pp. 111-135, 2004.

[3] Mahmood Akhtar "Comparison of Gene and exon prediction techniques for detection short coding regions," *International Journal of Information Technology* Vol. 11, No.8, 2005

[4] Sajid A.Marhon, Stefan C.Kremer "Gene prediction based on DNA spectral analysis: A literature survey," *Journal of computational biology*, vol. 18, no. 4, 2011.

[5] M.K.Hota, V.K.Srivastava, "DSP technique for gene and exon prediction taking Complex indicator sequence," *IEEE Signal Processing Magazine*, vol. 18, no. 4, pp. 8–20, 2009.

[6] Lun Huang, Mohammad Al Bataineh, G. E. Atkin, , Siyun Wang, Wei Zhang "A Novel Gene Detection Method Based on Period-3 Property," *31st Annual International Conference of the IEEE EMBS Minneapolis, USA*, pp.2-6, Sept 2009.

[7] Vikrant Tomar, Dipesh Gandhi, And Vijay Kumar "Digital Signal Processing for Gene Prediction," *IEEE Annual International conference*, pp.435-439, 2008.

[8] M.Roy, S. Barman "Spectral analysis of DNA sequence using Recursive Weiner Khinchine theorem- comparative approach" *IEEE International Conference On Recent Trend In Information Technology*, pp.315-318, 2011.

[9] Sajid A.Marhon, Stefan C.Kremer "Protein coding region prediction based on the adaptive representation method," *Canadian Conference on Electrical and Computer Engineering*, pp.415-418, 2011.

[10] Sylvain Robert Rivard, Jean-Gabriel Mailloux, Rachid Beguenane and Hung Tien Bui "Design of high performance parallelized gene predictors in Matlab" *BMC research notes*, vol.5, no. 4, pp.183-192, 2012.

[11] Benjamin Y. M. Kwan, Jennifer Y. Y. Kwan, Hon Keung Kwan "Spectral Techniques for classifying Short Exon and Intron Sequences" *International conference on biomedical engineering and biotechnology, pp.527-530*, 2012.

[12] S.Barman, M.Roy, S.Biswas, S. Saha "Prediction of cancer cell using digital signal processing," *International journal of engineering*, vol. 9, no. 3, pp.91-95, 2011.

[13] J.Setubal an J.Meidanis, "Introduction to computational molecular biology" CA, PWS Publishing Company, 1999.

[14] J. D. Watson and F. C. H. Crick, "Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid," *Nature*, vol. 171, pp. 737–738, 1953.

[15] R. J. Reece, "Analysis of genes and genomes," *England, John Wiley & Sons Ltd*, 2004.

[16] Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, "Essential cell biology," NY, Garland Publishing, 1998.

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