Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 5, May 2013, pg.332 – 337

RESEARCH ARTICLE

Genomic Signal Processing (GSP) Of Rheumatic Arthritis (RA) Using Different Indicator Sequences

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

[1] ramesh_k_b@hotmail.com; [2] prabhushankar.k.s@gmail.com

Abstract— DSP tools have been applied more prominently in the field of genomic signal processing for the prediction of genes in DNA sequences. By using a different indicator mapping and Fourier spectral characteristics, the method involves in the detection of period-3 periodicity of exons. In this paper we provide the experimental results in finding the gene using different indicator sequences by boosting the coding regions and suppressing the non-coding regions with application of fft.

Key Terms: - DNA; Rheumatic arthritis [RA]; Indicator sequences; Period-3 periodicity; Genomic signal processing [GSP] Technique

I. INTRODUCTION

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 (GSP)[2] 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 behaviour 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 rheumatic arthritis [RA] disease. The identification of the individual genes which are responsible for increased risk of disease will lead to a greater understanding of the actual mechanisms underlying disease. Although it has taken a considerable amount of effort to identify genes that increase the risk of disease, the hard work has only just begun! Much more work is needed to understand how variations in these genes alter the immune system and the inflammatory process. The hope is that this in turn should reveal new targets for treatment allowing patients' disease to be more effectively controlled. This work may also help in the identification of the environmental factors that increase the risk of disease. Individuals identified as being at risk because of their genes may then be able to avoid such environmental exposure, and thus prevent the disease occurring.

Rheumatoid arthritis is a chronic, autoimmune, inflammatory joint disease of unknown cause affecting approximately 1% of the population worldwide. Without early disease-modifying treatment, progressive and irreversible joint damage can occur that results in lifelong functional impairment. Rheumatoid arthritis can have a severe impact on health-related quality of life and socio-economic status and is associated with premature death. Recurring inflammation of the affected joints (i.e., arthritis) leads to a degradation of cartilage and to erosive destructions (erosions) of the bone. This affects physical function and mobility, and causes substantial short-term and long-term morbidity and an increased mortality rate compared to the unaffected population. It is a systemic disease which affects the joints as shown in below fig. 1in such a way that it can affect the whole body and internal organs such as lungs, heart and eyes. It affects approximately three times more women than men and onset is generally between 40-60 years of age although it can occur at any age. 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. Since abnormality of the DNA and coding regions are related to RA, we focus our attention to study the spectral characteristic of coding regions using digital signal processing.

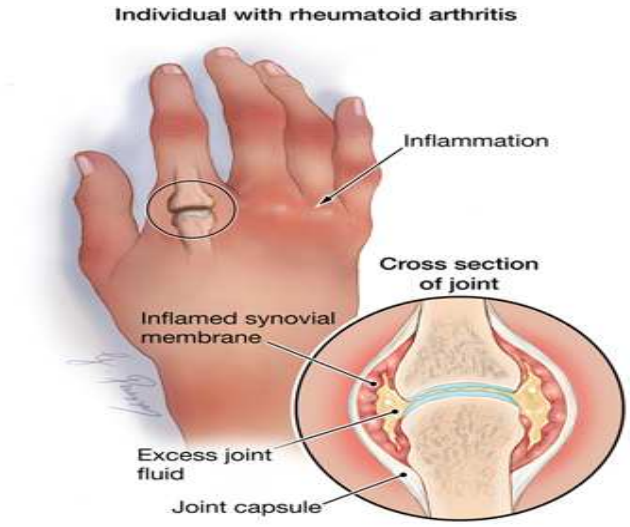

Fig. 1 Joints affected by rheumatic arthritis

Onset of the disease could be from one or more of the following factors. The role of genes plays an important part in contribution of RA. Cartilage becomes brittle with age and the body also loses its repairing capacities. Chances of arthritis increase with age. Being Overweight is a contributing factor as joint damage is partly dependent on the load, the joint has to support. Previous injury can cause damage to smooth working of a joint. Chances of Arthritis are associated with high stress jobs such as assembly line workers and heavy construction industry workers. Joint infections (septic joint), multiple episodes of gout, or certain other medical conditions, can lead to occurrence of arthritis. In this paper the authors have presented FFT power spectrum method to predict abnormality present in the nucleotide levels of a coding region using different indicator sequences.

II. OVERVIEW OF DNA

A DNA sequence is made from an alphabet of four elements, namely A, C, G, T (respectively adenine, cytosine, guanine, and thymine). The letters A, C, G, and T represent molecules called nucleotides or bases. Proteins are the building blocks of life and large numbers of functions are governed by proteins.

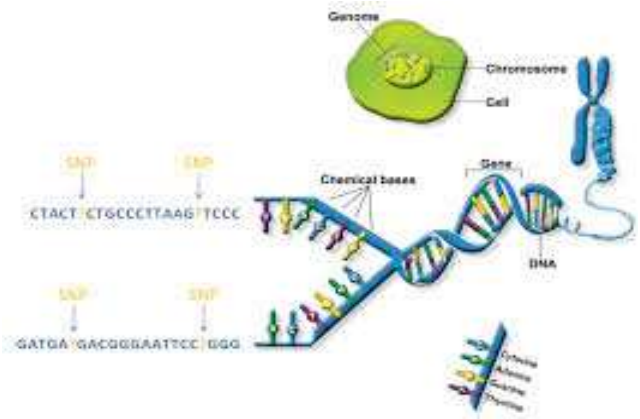

Fig. 2 Representation of DNA sequences and composition of chemical bases and genes

The DNA is found inside a special area of the cell called nucleus. Proteins are sequences made of amino acids. One of the most important goals in bioinformatics is to annotate biological sequences accurately and in a more productive way.

Protein-coding region prediction is an important step that has to be performed accurately before analysing and studying the proteins encoded by these regions of a DNA strand.

The genes present in the DNA sequences are divided into protein coding regions (exons) and intragenic spacers (introns) as shown in fig. 2. Genes are responsible for protein synthesis. Due to codon structure involved in the translation of base sequences into amino acids, period-3 component is present in the protein coding regions of DNA. Only the exons are involved in protein coding. Codons are a subsequence of three letters within the DNA sequence. The symbols in protein coding regions exhibit a triplet periodicity (TP) and this periodicity is rarely existed in non-coding regions [3].

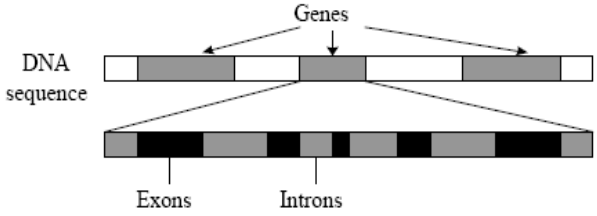

Fig. 3 Gene's structure

There are 64 possible codons (triplet). Each codon instructs the cell machinery to synthesize an amino acid. It is known that exons (or coding regions) are rich in nucleotides C and G whereas introns (or noncoding regions)[6] are rich in nucleotides A and T; and that protein coding regions of nucleotide sequences exhibit a period-3 property which is likely resulted from the three-base-length of codons used to generate amino acids.

III. DIFFERENT INDICATOR SEQUENCES

There exist different types of indicator sequences so that before applying transform on it, there is need to convert the nucleotides into numerical representation .Each strand of DNA consists of four nucleotides ,in which each base is converted into four signals.

Below table shows the different indicator sequences that can be used in performing the genomic signal proceeding.

	Name	**Numerical mapping**
1.	Single Galois Indicator	A = 0, C = 1, G = 3, T = 2
2.	Integer Number	A = 2, C = 1, G = 3, T = 0

3.	Real Number	A = -1.5, C = 0.5, G = -0.5, T = 1.5
4.	Complex Number	A= 1+j, C= -1-j, G= -1+j, T= 1-j
5.	Quaternary Code	A = 1, C = -j, G = -1, T = j
6.	Left-rotated Quaternary Code	A = j, C = 1, G = -j, T = -1
7.	EIIP	A=0.1260,C=0.1340,G=0.0806,T=0.1335
8.	Molecular Mass	A = 134, C = 110, G = 150, T = 125
9.	Atomic Number	A = 70, C = 58, G = 78, T = 66
10.	Paired Nucleotide Atomic Number	A or G = 62, C or T = 42
11.	Paired Numeric	A or T = 1, C or G = -1
12.	Complex Paired Numeric	A or T = j, C or G = -1

IV. METHODOLOGY

To perform the gene prediction analysis, DSP techniques have been implied to predict the period-3 property. Let $x(n)$ represent a discrete periodical signal, then its Fourier transform $X(k)$ is defined as

$X(K) = \sum_{n=1}^{N} x(n) e^{-2\pi jkn/N}$ k = 0, 1, 2….N-1

It is known that Period-3 property exist only in protein coding regions and can be exploited by to locate exons. Here we are interested in projecting the output by spectral analysis using

FFT[4] derived from DFT of the indicator sequences. It is possible to calculate the DFT more efficiently using Fast Fourier Transform (FFT) which reduces the number of operations.

Assuming for simplicity that N is a power of 2, N = 2^n,

by defining $W_n = e^{-2j\pi/N}$ and set M= N/2

The **overlap–add method (OA, OLA) [6]** is an efficient way to evaluate the discrete convolution of a very long signal x(n) with a finite impulse response (FIR) filter h(n)**:**

y[n] = x[n]* h[n] = y(n)$= \sum_{m=-\infty}^{\infty} h[m] * x[n-m]$

When sequence $x[n]$ is periodic, and N_x is the period, Using the windowing approach with a rectangular window length of L bases and an overlap width of L-3 bases between two adjacent windows, $y[n]$ is calculated, with the same period. In the region $M \le n \le N_x$, the resultant $y[n]$ sequence is correct. And if the next $M-1$ values are added to the first $M-1$ values, then the region $1 \le n \le N_x$ will represent the desired convolution.

Number of complex multiplications of the overlap-add methods are: $\left[\frac{N_x}{N-M+1}\right] N (log_2 N + 1)$.

V. RESULTS AND DISCUSSION

Figures below shows the results of exon prediction using different indicator sequences for the rheumatic arthritis of a sequence >gi|224586928|ref|NM_015967.4 taken from the ncbi and the simulation is done using hamming window with length of 300. Here the graph shows the distribution of frequency over power spectral density with respect indicator sequences indicated in table 1.With respect to indicator sequences the length and location changes and also the number of exons varies.

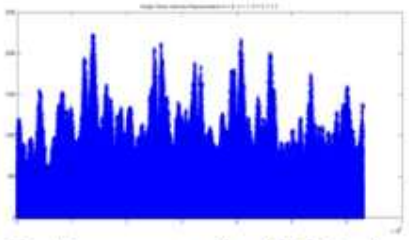

Fig 4 shows spectrum plot of RA for Single Galois Indicator with numerical mapping
A = 0, C = 1, G = 3, T = 2.

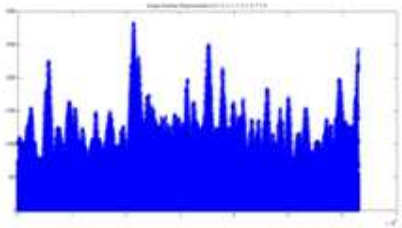

Fig 5 shows spectrum plot of RA for Integer Number with numerical mapping
A = 2, C = 1, G = 3, T = 0

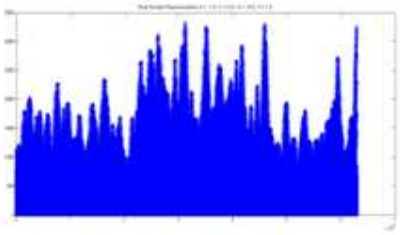

Fig 6 shows spectrum plot of RA for Real Number with numerical mapping
A = -1.5, C = 0.5, G = -0.5, T = 1.5.

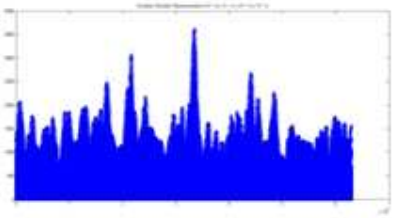

Fig 7 shows spectrum plot of RA for Complex Number With numerical mapping
A= 1+j, C= -1-j, G= -1+j, T= 1-j.

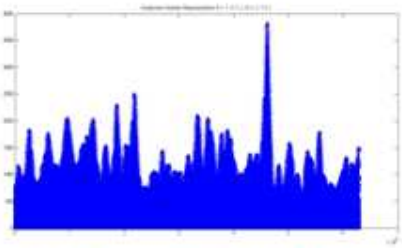

Fig 8 shows spectrum plot of RA for Quaternary Code with numerical mapping
A = 1, C = -j, G = -1, T = j.

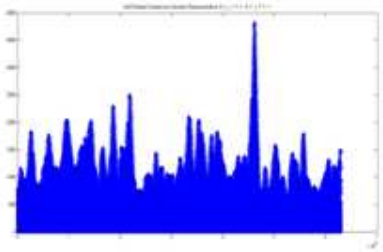

Fig 9 shows spectrum plot of RA for Left-rotated Quaternary Code with numerical mapping
A = j, C = 1, G = -j, T = -1

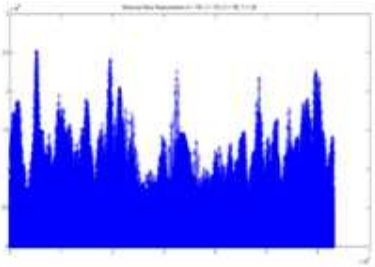

Fig 10 shows spectrum plot of RA for EIIP with numerical mapping
A=0.1260,C=0.1340,G=0.0806,T=0.1335

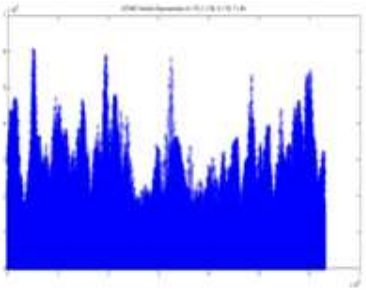

Fig 11 shows spectrum plot of RA for Molecular Mass with numerical mapping
A = 134, C = 110, G = 150, T = 125

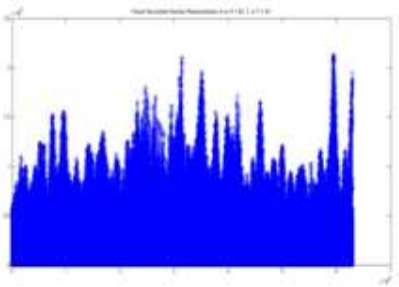

Fig 12 shows spectrum plot of RA for Atomic Number with numerical mapping A = 70, C = 58, G = 78, T = 66.

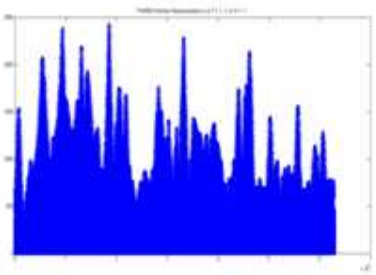

Fig 13 shows spectrum plot of RA for Paired Nucleotide Atomic Number with numerical mapping A or G = 62, C or T = 42

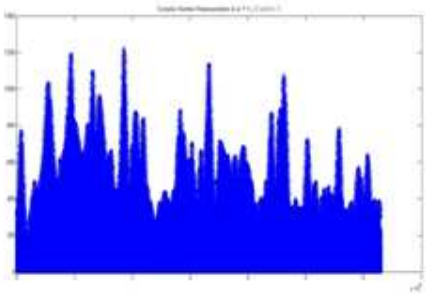

Fig 14 shows spectrum plot of RA for Paired Numeric with numerical mapping A or T = 1, C or G = -1.

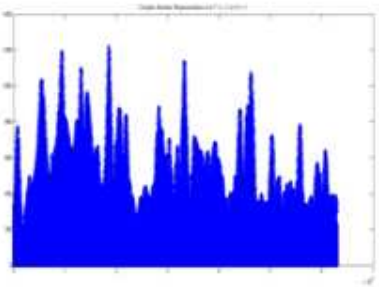

Fig 15 shows spectrum plot of RA for Complex Paired Numeric with numerical mapping A or T = j, C or G = -1.

VI. CONCLUSIONS

This paper presents the results obtained for different indicator sequences of rheumatic arthritis[RA] using the overlap and add algorithm of FFT method and results shows that different locations are obtained for the different indicator sequences and which can be useful in predicting where exactly the mutation takes place and exon lies in the sequence. The work will be useful for research community, expert doctors, specialists, which can benefit in drug design, drug discovery and pharmaceuticals and also common man will be hugely benefitted as there is awareness created among the population about the severity of the disease.

REFERENCES

[1] Petre G.Pop, Alin Voina "Numerical representations involved in DNA Repeats detection using spectral analysis" studies in informatics and control, vol 20, NO.2 June 2011.

[2] S.Barman , M.Roy, .S.Biswas, S.Saha "Prediction Of Cancer Cell Using Digital Signal Processing",Journal of Engineering 2011.

[3] Jennifer Yin Yee Kwan, Benjamin Yin Ming Kwan, Hon Keung Kwan " Spectral analysis of Numerical Exon and Intron Sequence" IEEE International Conference on Bioinformatics and Biomedicine Workshops , 2010.

[4] Benjamin Y. M. Kwan, Jennifer Y. Y. Kwan, Hon Keung Kwan "Spectral Techniques for Classifying Short Exon and Intron Sequences" IEEE ,pp : 219-226, 2012.

[5] Sajid A. Marhon And Stefan C. Kremer "Gene Prediction Based on DNA Spectral Analysis: A Literature Review" journal of computational biology Volume 18, Number 4, 2011.

[6] http://en.wikipedia.org/w/index.php?title=Overlap%E2%80%93add_method&oldid=541637833.

