MAT JOURNALS

Journal of Signal Processing
Volume 3 Issue 2

Detection of Rheumatic Arthritis Disease Based on Genomic Analysis by Applying Wavelet transform

Sunil S
Student (BMSPI)
Department of E & I E
R V College of Engineering, Bangalore

Dr.K B Ramesh
Head of Dept. & Associate Professor
Department of E &I E
R V College of Engineering, Bangalore

Dr. Vidya Niranjan
Professor & Associate Dean,Department of Biotechnology
R V College of Engineering, Bangalore

Abstract

In Recent years there is greater advance and innovations in bioinformatics. Bioinformatics is concerned with applying statistical and computational methods and also genomic signal processing techniques for analysis of data determined from sequenced DNA or RNA or Protein. To use genomic signal processing principle for analyze of DNA sequence first the DNA sequences of alphabetic string as to be converted into string of numeric sequence. This paper present the applications of wavelet transform based on the energy levels of approximation and detailed coefficients for sequence analysis with Chargaff's rule, internucleotide distance to compare two sequence similarities and determine the impact score so that to diagnose the Rheumatic Arthritis(RA).

Index Terms—*digital signal processing, DNA sequences, Chargaff's rule, genomes, internucleotide distance*

INTRODUCTION

DNA and proteomic sequence analysis is a major research topic in the fields of computer scientists, physicists, data analysis, statistical evaluations and mathematicians. Genomic Signal Processing is interesting area and has greater importance in the of science and engineering that has potential to evaluate evolution and differences in signal properties of DNAs and proteomics.DSP techniques provides compressive representation, different transformation techniques manipulation of digital signals for getting the information associated with the sequence. The signal may be Continuous or Digital signals which are represented in terms of sequences of numeric string in the case of the time series. [1] But in genomic sequences which is in the form of string of alphabets as represented by character strings symbols 4 alphabet sequences which are having the characters A, T, G and C (adenine, thymine, guanine, and cytosine) which represents the nucleotide bases of double stranded DNA. In case of proteomics, the alphabet size is twenty which are correspond to the possible amino acids. The amino acids are produced based in the DNA sequence in mRNA and tRNA.

In this paper,first an overview of essential concept for molecular biology is explained, second the DNA is mapped in to numeric sequence then, by applying the different comparison techniques for analyzing the DNA sequence of normal and abnormal rheumatoid arthritis patients. The applications of signal processing tools to study genomic sequences is done by applying multilevel signal decompositions using wavelet transform and compare the

energy's at each level of decompositions. Finally proposed Wavelet Transform for better analyzing of energy level of sequences.

BASIC CONCEPTS FROM MOLECULAR BIOLOGY

A. DNA

DNA is basically genetic information carrying element that is present in each and every cell. Single strands of DNA consists of smaller component many of that links as called nucleotide. The nucleotide is one of four possible amino acids (aa) namely Adenine (A), Cytosine (C) Thyamine (T) and Guanine (G). They can be represented as alphabet A, T, C, and G. Double stranded DNA has two distinct ends, that starts with 5'prime end on one side and the 3'prime end on other side. There 5'prime end of a nucleotides is linked to a 3'prime end of another nucleotides which is called as paring by strong single and double chemical bonds forming the long chain Double stranded DNA. Two single strands are complementary to each other i.e. a pairs to T and vice versa and C is pairs to G and vice versa. Bond formed has weak bonding but all together the bonds formed creates stable double helical structure. There two strands run in opposite direction to each other.

B. Proteins

Protein is also the one of the metabolic biomolecule consisting many building blocks of smaller component called amino acids. There are 20 possibility of amino acids that leads to the formation of proteins. The principal of Central dogma of biology states that DNAs codes for an RNAs and RNAs codes for proteins. The production of proteins takes place in two-stages, with RNAs that plays a key role in both stages. There stage one is called transcription there gene in there chromosomes of DNA is copied base by base into the RNA. RNA transcripts will be resulted there gene are then transported in the cell to the other molecular machine ribosome that plays task of translating RNAs to protein. Therefore DNA plays a very important role in central dogma of molecular biology. There character string is mathematically represents for further analysis and the length of the character strings represent a protein is relatively small in the hundreds while the other lengths of character string representing a DNAs in cell is typically in the millions and hundreds of millions.

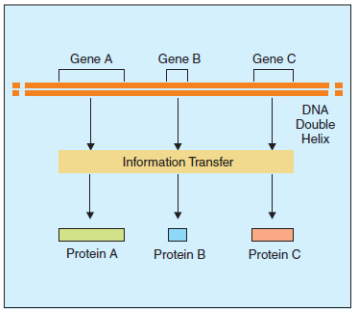

Fig 1. *Each gene contains the information to make a protein*

Protein molecules folds into 3-D structure which is complex forms the weak bonds with their own atoms they are functionally responsible to carry out all the essential functionalities in cell it is possible by binding to another biomolecules with required numbers of chemical bonds that connects neighboring atoms and other biomolecules.

D. Rheumatoid arthritis

Rheumatic Arthritis (RA) is a provocative illness which results in torment, inflammation, firmness, and decreases the capacity in the bone junctions or joints. These effects do happen when the resistant framework, which typically shields our body from attacking life forms, starts its assault against the film coating of the joints. The progression of RA can extend from mellow to serious. In rheumatoid arthritis the case where there body immunity system itself been attacked its

own healthy cells and tissue in the places of mainly joints. Later in chronic stages it attack internal organs of the human body system. Much of the time it is endless, which means it keeps going quite a while frequently a lifetime.

According to National Health Survey (NHS) that rheumatic arthritis affects almost 2-4% of world population and in India it is around 0.5 %, with women developing RA is more compared to men.

There is no cure for the rheumatoid arthritis patients but through physical and medications therapy can be useful to slow the disease's progression. The Mild cases of rheumatoid arthritis can be treated with an anti-inflammatory medications (NSAIDs). There are more severe cases which can be managed with the class of medications called antirheumatic drugs (DMARDS).

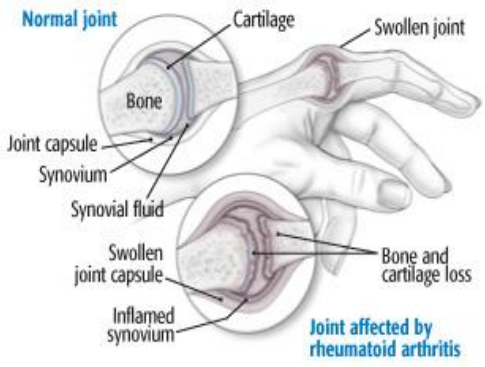

Fig 2. *Person affected with rheumatoid arthritis diseases.*

The Rheumatic Arthritis disease affected data and normal genome data is imported from GenBank. The proposed methodology is applied to several samples, and obtained results will provide the information about the presence or absence of disease and analysis of genes that are associated or in relation to the cause for RA disease.

METHODOLOGY

A. Chargaff's rule

Chargaff's rules is one that state that DNA from any of the living cell there should be 1:1 ratio which is also called as base Pair Rule of purine base and pyrimidine base more specially the quantity of guanine is should be equal to that of cytosine as well as the quantity of adenine is equal to thymine. Chargaff revealed two rules to facilitate that lead to the innovation of the twofold helix structure of DNA. So it has related that the adenine pairs in the company of the thymine and the cytosine pairs in the company of the guanine.

Adenine plus thymine is complementary to one another at the same time guanine plus cytosine are complementary. There has been two important reasons why DNA is composed of complementary bases. Chemical structure allows to form the strong hydrogen bonding with single and double between there bases that lead for the formation of the DNA.

B. The inter-nucleotoid distance of sequence

Suppose for DNA sequence S1 which has a length of n let their base nucleotide be x (x□ {A, G, C, T}) occurs for m times with positions of occurrence of their nucleotide be i_1, i2, i3….i_m. Let the equation d_x (m) =n-i_m is what the distance between there i_m th position of base to that of base x and the nth base position y in S1 sequence, we do not find the internucleotide distance between the same bases we calculate their distance between the bases values other bases values also depending on the number of input of maximum nucleotide distance [4]

C. Wavelet Transform

The wavelet transform has comparable that of Fourier transform is a completely different value function in signal processing. There most important

difference is that in Fourier transform decomposition of the given signal into the set of sines and cosines with different amplitude and frequencies in converse to that of the wavelet transform use functions which are is both localized in time and frequency.[5]The wavelet transform is expressed by the equation:

$$F(a,b) = \int_{-\infty}^{\infty} f(x)\, \psi^*_{(a,b)}(x)\, \mathrm{d}x \quad (1)$$

Where * represents the complex conjugate and function ψ is some function. The function has to be selected arbitrarily given that it follows certain rules.

The Wavelet transform is fact that infinite sets of the several transforms depend on the merit function that is used for their computation. The main reason behind why that we hear term "wavelet transform" in very special situation is because of its applications. Where it is based on the wavelet orthogonality. We use orthogonal wavelets for the discrete wavelet transform and for continuous wavelet transform we use the non-orthogonal wavelets for their applications.

These DWT and CWT transforms has to follows the following properties. Discrete wavelet transform return the data vector as the same length as that of input. Even has this vector many data that are nearly zero. This results to fact that it well decomposes into a set of wavelet. They are orthogonal which are translations and scaling. There by we get the decomposed signal to a same or to the lower number wavelet coefficient spectrum as of number of signal data point. This type of wavelet spectrum is very good for signal processing and compression also for to get no redundant information's data. Discrete wavelet transform (DWT) is the implementation of wavelet transforms by using discrete sets of wavelet scales and translations. discrete wavelet transform decomposes the signal into the mutually orthogonal set of wavelets this is the main dissimilarity of that of continuous wavelet transform and the CWT implementation for discrete time series data set is termed as discrete-time continuous wavelet transform (DT-CWT).

The wavelet is been constructed from the scaling function where it follows all the it's scaling properties. The main restriction for this scaling functions is that it must have orthogonal to the discrete translations that implies the some mathematical conditions on their scaling which it is mentioned as dilation equation

$$\phi(x) = \sum_{k=-\infty}^{\infty} a_k \phi(Sx - k) \quad (2)$$

Here *S* is known as scaling factor. The area between there functions has to be normalized and the scaling function as to be orthogonal to that of the integer translations as show in equation

$$\int_{-\infty}^{\infty} \phi(x)\, \phi(x+l)\, \mathrm{d}x = \delta_{0,l} \quad (3)$$

After following conditions we obtained result of these respective equations that is there finite set of coefficients approximate coefficients a_k which will be the define scaling function and wavelet. Wavelet has been obtained from the scaling function of length *N* where this is integer.

The decompose of signal use the orthonormal basis set of wavelets. That simplifies calculation because there are only few of approximate coefficients a_k which are nonzero.

There is a different methods and types implementation DWT algorithm. One of the most known widely used one is Mallat (pyramidal) algorithm. This algorithm use two of the filters one of which is smoothing filter and other one is non-smoothing filter which is constructed from wavelet coefficients of their filters which

is generally used to fetch data's used from all there scales. The total length number of data is $D = 2^N$ used and the signal length is L, then first $D/2$ data scaled at $L/2^{N-1}$ which are computed, then $(D/2)/2$ data scale at $L/2^{N-2}$ is up to obtain finally two dataset at scale of $L/2$. The obtained results of the algorithm is array of the same length the same as input one where data are usually sorted from the largest scales to the smallest ones.

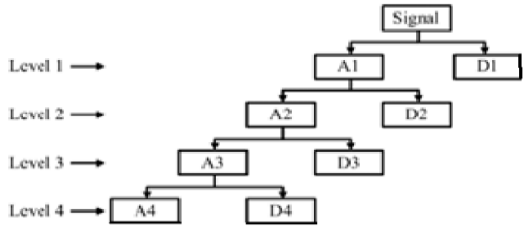

Fig 3. *Multilevel wavelet decomposition. (Level 4)*

We first demonstrate our method in the two-phase case. In the first level the flat file is downloaded from the database the flat file may consist of sequence Nucleotide, DNA, RNA, Protein sequences. From the database the person affected with rheumatoid arthritis and normal human genome is taken and wavelet transform [5] is applied to the two sequence and the scalogram of two sequence is obtained which shows the visual changes between the normal person with that of diseases affected person

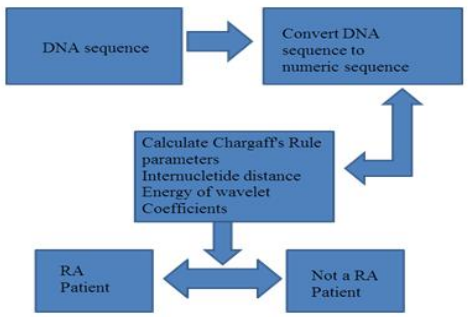

Fig 4. *Block diagram for steps followed for genomic analysis*

RESULTS

Table.1 *Abnormal sequence Chargaff's parameters*

A/T	G/C	AT/GC
1.1023	1.2112	1.0393
1.1327	1.2755	1.0807
1.0472	1.3158	1.1818
1.1795	1.2273	1.0408

Table.2 *Normal sequence Chargaff's parameters*

A/T	G/C	AT/GC
1.3959	0.8876	1.0446
1.7697	1.0106	1.2026
1.4431	0.9387	1.1314
1.1101	1.3488	1.1386

A total of 20 sequences has been tested for Chargaff's rule eight of which is show in the table. The impact sore from the table is for a sequence to be normal the value of A/T should lay within 1.3959 to 1.7697 and the value of G/C should lay within 0.8876 to 1.3488.

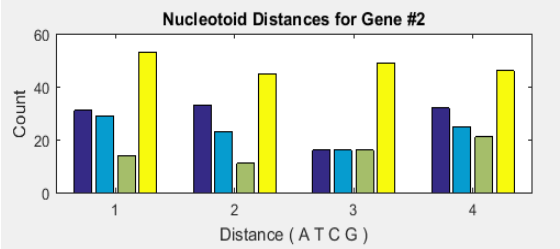

Fig 5: *Internucletide distance for maximum distance of four for abnormal sequence.*

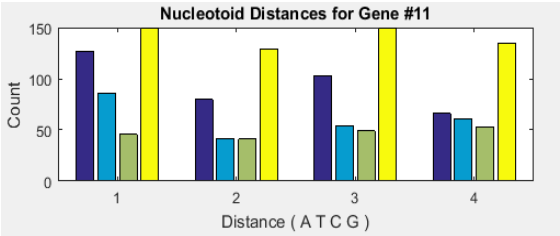

Fig 6: *Internucletide distance for maximum distance of four for Normal sequence.*

For the maximum Internucletide distance of four the occurrences of DNA bases in both normal and abnormal sequences is show in the Fig 5 and Fig 6. We find from

the figures that there abnormal sequence has less nucleotide bases when compared to the normal sequences.

The DNA that is converted to numeric sequence data is applied with wavelet decomposition[6] of four level using symlets 4 wavelet,[7]There energy levels are calculated. The energy 80 to 82 of approximate coefficients of the impact score.

Table3: *Energy level variation between there wavelet coefficients for the normal sequences.*

A4	80 to 82
D3	8.4 to 9.6
D2	4.6 to 4.8
D1	1.1 to 1.4

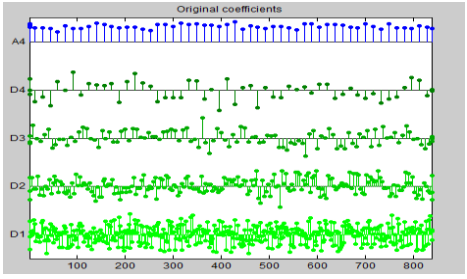

Fig 7: *Level 4 decomposition of wavelet coefficients for normal sequence.*

CONCLUSION

In this paper Chargaff's rule, Internucletide distance and the energy of wavelet coefficients are used for differentiating the normal and abnormal sequence and determined the impact score that effect the abnormal sequence that can be used by the doctor for their precise diagnosis, differentiation from patients with RA and non-RA patients.

FUTURE SCOPE

The percentage of energy with each coefficient which represents the active region in the genomic sequences can be represented by scalogram. There scalogram give the active proteins that are present in there sequences can be used as future study.

REFERENCES

1. K. Deergha Rao, Senior Member, IEEE, and M. N. S. Swamy, Life Fellow, IEEE "Analysis of Genomics and Proteomics Using DSP Techniques", IEEE transactions on circuits and systems—i: regular papers, vol. 55, no. 1, february 2008.
2. P.Saranya,V.Harigopalkrishna,D.Murali,M.Ravikumar,M.Sujatha" Analysis of Genomic and Proteomic Sequence Using Fir Filter" Journal Of Modern Engineering Research,IJMER,ISSN: 2249–6645, Vol. 4, Iss. 2 ,Feb. 2014 |105|
3. Hamidreza Saberkari, Mousa Shamsi, MohammadHossein Sedaaghi, Faegheh Golabi "Prediction of protein coding regions in DNA sequences using signal processing methods"IEEE Symposium on industrial and applicaiona(ISIEA) 978-1-4673-3005-3/12/$31.00 ©2011
4. Yushuang Li, Yanfen Lv, Xiaonan Li, Wenli Xiao, Chun Li" Sequence comparison and essential gene identification with new inter-nucleotide distance sequences" journal of Theoretical Biology, DOI: 10.1016/j.jtbi.2017.01.031
5. Shiwani Saini,Lillie Dewan" Application of discrete wavelet transform for analysis of genomic sequences of Mycobacterium tuberculosis" SpringerPlus (2016) 5:64, DOI 10.1186/s40064-016-1668-9
6. T. M. Inbamalar,R. Sivakumar" Improved Algorithm for Analysis of DNA Sequences Using Multiresolution Transformation" Hindawi Publishing Corporation The Scientific World Journal Volume 2015, Article ID 786497
7. J.A. Tenreiro Machado, António C. Costa,Maria Dulce Quelhas" Wavelet analysis of human DNA" 0888-7543 © 2011 Elsevier Inc. All rights reserved. doi:10.1016/j.ygeno.2011.05.01