# Susceptibility Gene Prediction in Hereditary Disease Retinoblastoma

S. Sumathi[1], Dr. R. Dhaya[2], Dr. R. Kanthavel[3]

[1]Research Scholar, Velammal Engineering college, Chennai, India
[2]Associate professor, Rajalakshmi Engineering College, Chennai, India
[3]Vice Principal, Rajalakshmi, Institute of Technology, Chennai, India

***Abstract***— *Nowadays Bioinformatics, proteomics and Genomics are the most intriguing sciences to understand the human genome and diseases. Several hereditary genetic diseases like Retinoblastoma involve a sequence of complex interactions between multiple biological processes. With this paper, genetic similarities were found within a selected group of patient's DNA sequences through the use of signal processing tools. DNA, RNA and protein sequences have similarities in structure and function of the gene with their location. In this paper, we introduce a novel method using scoring matrix and wavelet windowing, for the integrative gene prediction. The proposed methods not only integrate multiple genomic data but can be used to predict gene location, gene mutation and genetic disorder from the multi-block genomic data. The performance was assessed by simulation.*

***Keywords— Gene, scoring matrix, WWM.***

## I. INTRODUCTION

Retinoblastoma is a malignant cancer of the increasing retinal cells caused in the majority cases by mutations in both copies of the RB1 gene. The RB1 gene is a tumor suppressor gene, located on the genetic material, chromosome 13q14 and is the first cloned human cancer gene. The gene codes for the tumor suppressor protein pRB, which by binding to the transcription factor E2F, inhibits the cell from entering the S-phase during mitosis. Latest facts about retinoblastoma suggests that post-mitotic cone precursors are uniquely sensitive to pRB depletion and may be the cells in which retinoblastoma originates. The occurrence and viability of retinoblastic cells may be more complex than suggested by simple loss of function of the RB1 alleles. Hereditary retinoblastoma demonstrate close relation of the gene for this cancer with genetic locus for esterase D. Data are presented here in support of the hypothesis that at least one disease, the retinoblastoma observed in children is caused by two mutational events.

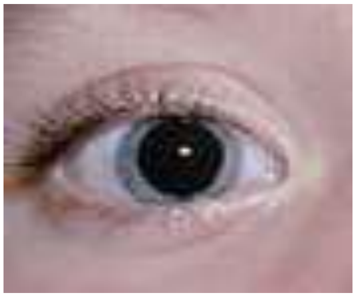


*Fig.1: Healthy eye*

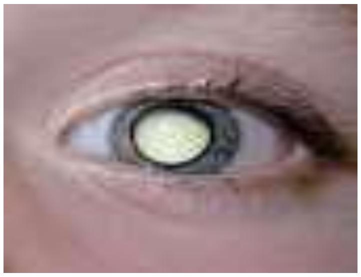


*Fig.2: Retinoblastoma affected eye*

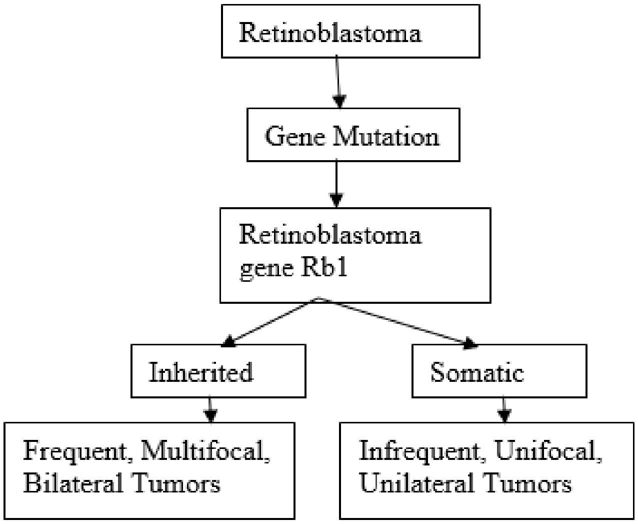

Retinoblastoma
Gene Mutation
Retinoblastoma gene Rb1
Inherited
Somatic
Frequent, Multifocal, Bilateral Tumors
Infrequent, Unifocal, Unilateral Tumors

*Fig.3: Flow graph of gene mutation in Eye*

In 5% of retinoblastoma cases with germline mutations the ancestor history is positive. The risk for developing bilateral and multifocal retinoblastoma is high and the age of inception is around 15 months. The mean number of tumors is about 5 in the two eyes. The offspring of a parent with bilateral retinoblastoma have a 50% probability of developing a tumor and 50% possibility of inheriting the germline mutant allele. Reduced reentrance of 10 to 15% lowers the estimated occurrence of disease from 50% to 25%. Individuals who have mutations in both alleles somatically do not have a mutation in their germ cells and therefore usually transfer no tumor risk to their offspring.

## II. DNA SEQUENCE

Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) are consisting of a nucleobases, a pentose sugar and a phosphate group. DNA nucleobases are Cytosine (C), Guanine (G), Adenine (A) and Thymine (T) and RNA nucleobases are Cytosine (C), Guanine (G), Adenine (A) and Uracil (U)[1]. In recent years huge databases available for genetic information as open source which lead to a huge progress in bioinformatics; if a genetic sequences are known then this information could be a very important in early disease diagnosis, drug discovery for it.[2] It leads to Biological sequence alignment a field of Bioinformatics and Computational Biology. It's aim analyzing similarities between DNA, RNA or protein sequences, to predict the genetic relationship between organisms and structural or functional relationships.

Each segment in DNA is called a gene. Genes control the protein synthesis and regulate most of the activities inside a living organism. All the genetic information is copied when a cell divides. When a change occurs in the base sequence of a DNA strand, it is called a mutation. These mutations can lead to diseases or the death of a cell.

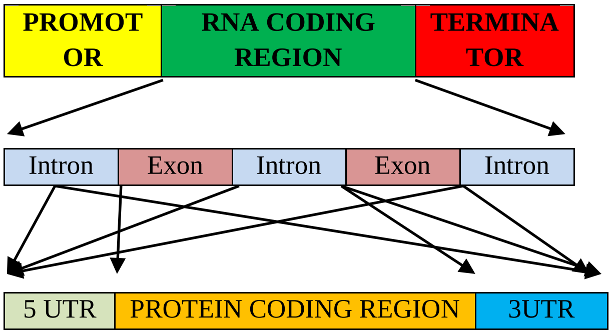


*Fig.4: Gene structure*

The numerical representation of DNA sequences becomes very essential as almost all DSP techniques require two parts: mapping the symbolic sequence into a numeric and calculating a kind of transform of the resultant numeric series [2]. Most of the numerical representations associate one numerical value to one position in the sequence using numerical values related to each nucleotide and, finally, reveal the existence or the nonexistence of a certain nucleotide in a specific position [3]. Another approach could be to include information about the number and type of repeated nucleotides to generate only one numerical value for each DNA subsequence which may be associated with a recur. This representation needs a mapping algorithm which use distances to determine similar subsequences and then evaluate a consensus sequence for these subsequences to generate candidates.

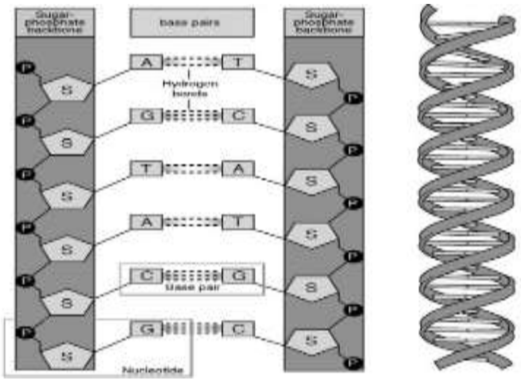
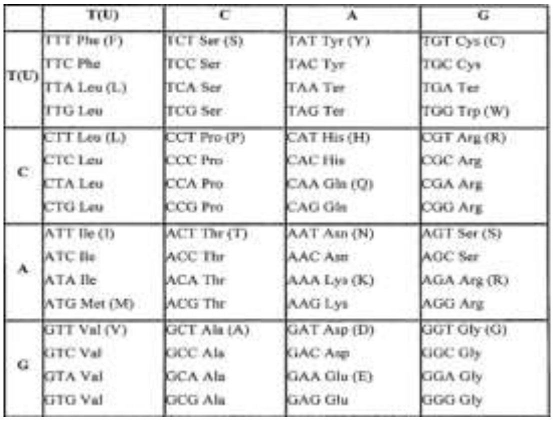


*Fig.5: DNA Helical Structure*

## III. GENE PREDICTION

Gene Prediction refers to detect the locations of the protein-coding regions of genes in a lengthy DNA sequence. Signal processing techniques offer a huge guarantee in analyzing genomic data because of its digital nature. Signal processing analysis of bio-molecular sequences is stalled by their representation as strings of alphabet characters.

*Table.1: Genetic code*



| | T(U) | C | A | G |
|---|---|---|---|---|
| T(U) | TTT Phe (F) | TCT Ser (S) | TAT Tyr (Y) | TGT Cys (C) |
| | TTC Phe | TCC Ser | TAC Tyr | TGC Cys |
| | TTA Leu (L) | TCA Ser | TAA Ter | TGA Ter |
| | TTG Leu | TCG Ser | TAG Ter | TGG Trp (W) |
| C | CTT Leu (L) | CCT Pro (P) | CAT His (H) | CGT Arg (R) |
| | CTC Leu | CCC Pro | CAC His | CGC Arg |
| | CTA Leu | CCA Pro | CAA Gln (Q) | CGA Arg |
| | CTG Leu | CCG Pro | CAG Gln | CGG Arg |
| A | ATT Ile (I) | ACT Thr (T) | AAT Asn (N) | AGT Ser (S) |
| | ATC Ile | ACC Thr | AAC Asn | AGC Ser |
| | ATA Ile | ACA Thr | AAA Lys (K) | AGA Arg (R) |
| | ATG Met (M) | ACG Thr | AAG Lys | AGG Arg |
| G | GTT Val (V) | GCT Ala (A) | GAT Asp (D) | GGT Gly (G) |
| | GTC Val | GCC Ala | GAC Asp | GGC Gly |
| | GTA Val | GCA Ala | GAA Glu (E) | GGA Gly |
| | GTG Val | GCG Ala | GAG Glu | GGG Gly |

## IV. NUMERICAL REPRESENTATION

The arithmetical depiction of a DNA sequence is given as a chain of integers derived from a unique graphical representation of the regular hereditary code. This numerical representation is appropriate for the quantitative analysis of the sequences.

### 4.1 LD matrix

LD matrix is used to calculate linkage disequilibrium values."composite" for LD composite measure, "r" for R coefficient (by EM algorithm), "dprime" for D', and "corr" for correlation coefficient. The method "corr" is equivalent to "composite", when SNP genotypes are coded as: 0 – BB, 1 – AB, 2 – AA. Matrix elements adjacent to the main diagonal represent the extent of the line segments producing the line.

### 4.2 Transition matrix

Transition matrix is used for transitions from one kind of base to another. For a given DNA sequence ‘s’ it can construct a 4×4 matrix A = (tij), where tij means the number of times a given base being succeeded by another in the sequence. A is called the transition frequency matrix of s. We can construct a matrix P = (Pij) by dividing each element by the total of all entries in A. Such a matrix represents the relative frequency of all the possible types of transitions, and is called the transition proportion matrix of s. The initial mapping of DNA to binary which represents DNA with four binary indicator sequences showing the presence ‘1’ and absence ‘0’ of the relevant nucleotides at locations 'n'.

### 4.3 Complex representation

The complex representation is based on the assumption that coefficients of the four 3-D tetrahedron vectors representing each DNA letter are either +1 or -1. The dimensionality of the resultant bipolar representation can be condensed to two.

## V. DSP TECHNIQUES

### A. DFT

Fourier transform is used to detect the likely coding regions in DNA sequences, by computing the amplitude profile of this spectral component which is a sharp peak at frequency f = 1/3 in the power spectrum. The strength of the peak depends obviously on the repetition of gene. This gives relatively good results but it is dependent on DNA sequence and thus requires computation before processing of the mapping scheme for gene prediction. The DNA sequence to be generated from a white random process through an all pole system and thus used Auto-Regressive modeling to replace Fourier analysis for exon prediction.

### B.STFT

In non-stationary signals, The Short Time Fourier Transform (STFT) is an algorithm frequently used for the DFT-based spectral analysis. In the STFT, the time signal is divided into short segments and a DFT is calculated for each one of these segments. Spectrogram, a three dimensional graph called is obtained by plotting the squared magnitude of the DFT coefficients as a function of time.

### C. DWT

The Discrete Wavelet Transform is a mathematical tool that can be used very effectively for non-stationary signal analysis. The DWT, for which an algorithm called Fast Wavelet Transforms (FWT) allows a very efficient calculation. Methods based on a modified Gabor-wavelet transform (MGWT) for the identification of protein coding regions also exists.

## VI. WAVELET WINDOW METHOD

A Wavelet Transform Modulus Maxima (WTMM) is defined as a point ($x_0$, $t_0$) such that

$$lW_{x0,\, t}l \leq lW_{x0,\, t0}l$$

when *t* belongs to either a right or the left neighbourhood of $t_0$, and

$$lW_{x0,\, t}l \leq lW_{x0,\, t0}l$$

when *t* belongs to the other side of the neighbourhood of $t_0$. We describe maxima line, any connected curve in the scale space (x, *t*) along which all points are WTMM.

## VII. RESULTS AND DISCUSSIONS

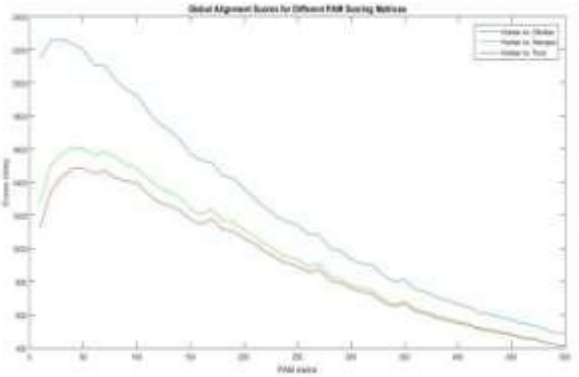


*Fig.1: Best of DNA seq*

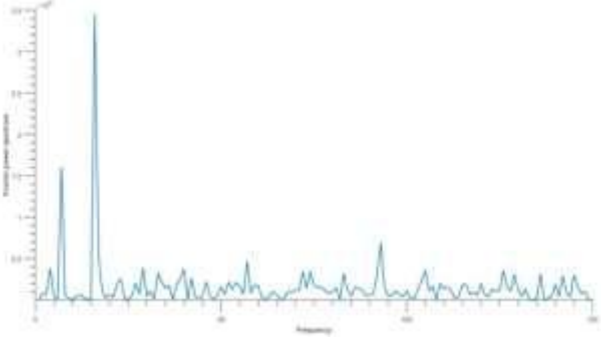


*Fig.2: Fourier spectrum of DNA sequence*

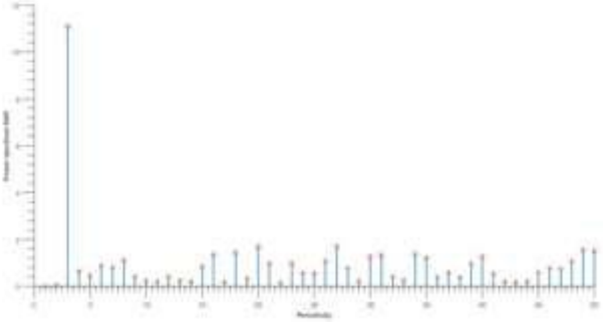


*Fig.3: Power spectrum SNR of DNA sequence using WWM*

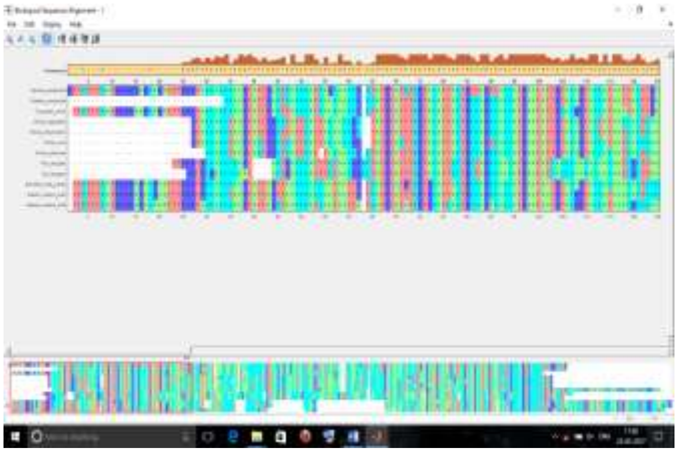


*Fig.4: Spectrum analysis of normal and abnormal DNA sequence*

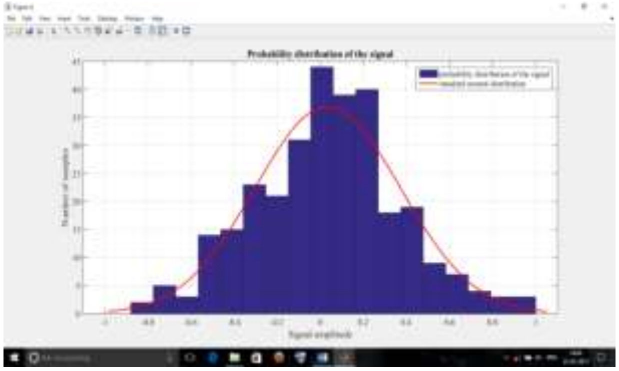


*Fig.5: PDF of given sequence*

## VIII. CONCLUSION

In this work, a new analyzing wavelet window and scoring matrix method for the prediction of protein coding regions has been proposed. The wavelet window method can be applied to predict different coding regions of different lengths. The selection of the value of the window length has always been a problem in DSP based methods as it has an effect on the gene prediction. Future work can focus on integrating this technique to refine the predicted location of gene and protein coding regions.

## REFERENCES

[1] Asmaa G.Seliem, Wael Abou El-Wafa, etl., Parallel Smith-Waterman algorithm hardware implementation for ancestors and offspring gene tracer

[2] Lim KG, Kwoh CK, Hsu LY, Wirawan A., "Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance", Brief Bioinformatics, 2012, vol. 14, no. 1, pp:67-81.

[3] Lorenzo-Ginori,J.V.,Rodriguez-Fuentes,A.,Abalo,R.G.,Rodrigues,R.S., "Digital signal processing in the analysis of genomic sequences", Curr. Bioinformatics,2009, 4, pp:28-40.

[4] Pop, P.G., Voina,A., "Representations Involved in DNA Repeats Detection Using Spectral Analysis", Studies in Informatics and Control, 2011, vol. 20, no. 2, pp:163-180.

[5] V.I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals", Soviet Physics Doklady, 1966, 10, pp:707-10.

[6] Deza, M.M., Deza E., Encyclopedia of Distances, Springer, 2009.

[7] NCBI at http://www.ncbi.nlm.nih.gov/genban/

[8] Alkan C, Ventura M, Archidiacono N, Rocchi M, Sahinalp SC, Eichler EE., "Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data", PLoS Comput Biol. 2007, Sep; 3(9),pp:1807-18.

[9] D. Anastassiou, "Genomic signal processing," IEEE Signal Processing Mag., vol. 18, pp. 8–20, 2001.

[10] D. Cohen, I. Chumakov, and J. Weissenbach. A first-generation physical map of the human genome. Nature, 698–701, Dec 1993.

[11] T. Lengaeur, editor. Bioinformatics - From Genome to Drugs, volume II: Applications of Methods and Principles in Medicinal Chemistry. Wiley-VCH Verlag, Weinheim, 2002.

[12] K. Usdin The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. Genome Res;18:1011–9, 2008.

[13] F. Glutamine, repeats and neurodegenerative diseases: molecular aspects. Trends in biochemical sciences, 24(2), 58-63, 1999.

[14] V. R. Chechetkin and A. Y. Turygin, ‘‘Search of hidden periodicities in DNA sequences,’’ J. Theoretical Biol., vol. 175, pp. 477---497, 1995.

[15] M. Buchner and S. Janjarasjitt, “Detection and visualization of tandem repeats in DNA sequences,” IEEE Trans. Signal Process., vol. 51, no. 9,pp. 2280–2287, Sep. 2003.

[16] D. Sussillo, A. Kundaje, and D. Anastassiou, “Spectrogram analysis of genomes,” EURASIP J. Appl. Signal Process., vol. 2004, no. 1, pp. 29–42,2004.

[17] Zhou Zhi-min, Chen Zhong-wen, “Dynamic Programming for Protein Sequence Alignment”, *International Journal of Bio-Science and Bio-Technology*, Vol. 5, No. 2, pp. 141–150, April, 2013.

[18] S. S. Ray, S. Ghosh, and R. Prasad, “Low-cost hierarchical memorybased pipelined architecture for DNA sequence matching”, *IEEE INDICON 2014*, pp. 1–6, 11-13 Dec. 2014.

[19] S. Ghosh, S. Mandal and S. Saha Ray, “A scalable high-throughput pipeline architecture for DNA sequence alignment”, *IEEE TENCON 2015 IEEE Region 10 Conference*, Macao, pp. 1–6, Nov. 2015.

[20] P. K Lala and J. Parkerson,“A CAM (Content Addressable Memory)- based architecture for molecular sequence matching”, International Conference on Bioinformatics and Computational Biology, Las Vegas, July 18–21, 2011.