PREDICTION OF INDIVIDUAL GENETIC RISK

1Ms. A.Priya Dharshini, 2 Ms. S.Sangeetha, 3 Ms. R.Meena

1, 2 Student, 3Professor
1, 2, 3 Computer Science and Engineering,
1, 2, 3 Prathyusha Engineering College, India.

***Abstract*-**Our paper is subjected to Coronary heart diseases (CHD), which is a leading cause of mortality and morbidity worldwide due to the lifestyle of an individual person there may be changes in genetic risk factors. To calculate GC values for specific heart disease or DNA sequence, where GC-content (or guanine-cytosine content) is the percentage of nitrogenous bases on a DNA or RNA molecule. In order to avoid redundancy identical sequences are merged, regardless of whether they are from the same or different species. Each sequence is given a stable and unique identifier (UPI), making it possible to identify the same protein from different source databases. Since there is a correlation between the single sequences, which is completely overcome with Uniprot database. With the help of single sequence, the mutation can be identified later. Additionally, to read and make a multiple alignment of the protein sequences from the Fasta file for both long and short DNA sequences, the CLUSTAL software is used. Apart from Clustal software, we are also including the protein gap, where it is used to estimate the divergence between two sequences, and it's usually measured in quantity of evolutionary, thereby calculating the genetic distances between DNA (or mRNA) sequences and to build a phylogenetic tree based on the distance matrix.

***Keywords*:GC content, Clustal, FASTA, Protein gap, Uniprot, DNA, RNA**

INTRODUCTION

GC-content (or guanine-cytosine content) is the percentage of nitrogenous bases on a DNA or RNA molecule that are either guanine or cytosine (from a possibility of four different ones, also including adenine and thymine in DNA and adenine and uracil in RNA). GC content is usually expressed as a percentage value, but sometimes as a ratio (called G+C ratio or GC-ratio). GC-content percentage is

Calculated as,

$$\frac{G+C}{A+T+G+C}$$

Whereas the AT/GC ratio is calculated as,

$$\frac{A+T}{G+C}$$

A protein gap is used to estimate the divergence between two sequences, and its usually measured in quantity of evolutionary, thereby calculating the genetic distances between DNA (or mRNA) sequences and

to build a phylogenetic tree based on the distance matrix. And now, a mutation is a change that occurs permanently in the DNA sequence thatmakes up a gene which differs from most people. Mutations ranges in size, it starts from a single DNA strand to larger chromosomes in turn to multiple genes. Mutation occurs because of four classes and they are (i)Spontaneous mutations, (ii)Naturallyoccurring DNA damage (most common factor to cause mutation in an individual orat the time of birth), (iii)Errors caused due to DNA repair, (iv)Induced mutations (artificial) by mutagens (a chemical or physical agent that causes genetic mutation in an individual).

EXISTING SYSTEM

This system uses GC (or Guanine-Cytosine content) and GC3 (Proportion of G and C in the third position of the codons) content of the protein coding sequences (CDS) of 22 metazoan and non-metazoan species using the available data from the genome databases are investigated. And then, they have compared the distribution of the GC and GC3 contents of the CDS of the vertebrate species and non-bilaterian animals. The mean of the GC content of the non-bilaterian *T. adhaerens* CDS is 38%, and thus the lowest seen in any metazoan genome so far investigated. For normalization of gene expression levels the geometric mean of the most stable reference genes was used. Finally, it confirms the narrow range of the GC3 content of *T. adhaerens* CDS and emphasizes the differences to other bilaterian and non-bilaterian metazoan species.

DISADVANTAGES OF EXISTING SYSTEM

- Predicted only with the calculation of GC values.
- They have analyzed accuracy of 50% only and it is uncertain to do the micro level study of data.

PROPOSED SYSTEM

In proposed system, Genetic level variation is compared for various heart disease types using the ATGC sequence. The DNA sequence is analyzed for the prediction of GC content variation that is to reveal the cases of horizontal transfer or reveal biases in mutation and then making an alignment through the clustal software, which is used to determine the protein gap between the two DNA sequences, which is done for different types of heart disease sequence. And the observation is made which leads to the next step of the study which includes the formation of phylogenetic tree, where it is used to estimate the divergence between two sequences, and its usually measured in quantity of evolutionary, thereby calculating the genetic distances between DNA (or mRNA) sequences that are read from the Uniprotdatabase. Then a protein tree is plotted for both rooted and unrooted types of heart disease genome cell for the individuals.

ADVANTAGES OF PROPOSED SYSTEM

- Large amount of genome sets is to be read and analyzed while compared to the existing system.
- The graphical representation makes better understanding of the result analysis.
- Identifying mutation with the help of protein tree.
- This improves accuracy of coronary heart disease for both men and women.

METHODOLOGY USED

There are certain important modules used for these analyses they are as follows:

a) Reading FASTA file and Analysis of gene

b) Counting of the Nucleotides

c) Calculation of GC value

d) Formation of phylogenetic tree

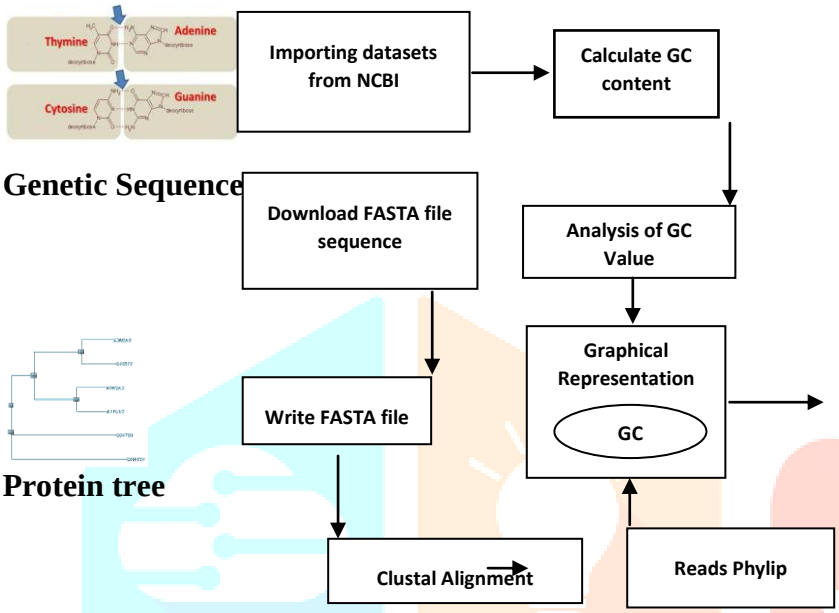


Figure.1: The work flow of the project

Reading FASTA file and Analysis of gene

In bioinformatics, FASTA format is a text–based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes. The FASTA file is downloaded from NCBI (National Center for Biotechnology Information). The most common database is GenBank which can be accessed using the code (GB | accession | locus).

The format also allows for sequence names and comments to precede the sequences. The FASTA file is used to parse sequences using text- processing tools and scripting languages like python, Ruby and Pearl. The FASTA file can be downloaded with free distribution of FASTA (see fasta 20. doc, fasta VN.doc or fastaVN. me - here VN is a Version Number).

A sequence in FASTA is represented as a series of lines, that is it should be no longer than 120 characters and do not exceed 80 characters. The first line in a FASTA file starts with a “>” symbol or “;” symbol (less frequently used). The FASTA sequence commonly ends with an asterisk (*) character to leave a blank line between the description and the sequence.

a. Sequence representation

After the header line and comments, one or more lines describing the sequence, and each line should be fewer than 80 characters. Sequences may be protein sequences or nucleic acid sequences. The nucleic acid codes and the amino acid sequences can be represented using characters (i.e. lower case and upper case but notintegers).

Some of the nucleic acid codes supported are

A adenosine	C cytidine	G guanine
T thymidine	N A/G/C/T (any)	U uridine
K G/T (keto)	S G/C (strong)	Y T/C (pyrimidine)
M A/C (amino)	W A/T (weak)	R G/A (purine)
B G/T/C	D G/A/T	H A/C/T
V G/C/A	- gap of indeterminate length	

Figure.2: Nucleic acid codes

Some of the amino acid codes are

A	alanine	P	proline
B	aspartate/asparagine	Q	glutamine
C	cystine	R	arginine
D	aspartate	S	serine
E	glutamate	T	threonine
F	phenylalanine	U	selenocysteine
G	glycine	V	valine
H	histidine	W	tryptophan
I	isoleucine	Y	tyrosine
K	lysine	Z	glutamate/glutamine
L	leucine	X	any
M	methionine	*	translation stop
N	asparagine	-	gap of indeterminate length

Figure.3:Amino acid codes

b. Example of a FASTA file format

A FASTA file can contain more than one sequence. If a FASTA file is of sequences, then for each sequence it will have a header line starting by the sequence itself.

Example FASTA file

```
>mysequence1
ACATGAGACAGACAGACCCCCAGAGACAGACCCCTAGACACAGAGAGAG
TATGCAGGACAGGGTTTTTGCCCAGGGTGGCAGTATG
>mysequence2
AGGATTGAGGTATGGGTATGTTCCCGATTGAGTAGCCAGTATGAGCCAG
AGTTTTTTACAAGTATTTTTCCCAGTAGCCAGAGAGAGAGTCACCCAGT
ACAGAGAGC
```

Figure.5: FASTA file format

Counting of the Nucleotides

The nucleotide database is the collection of sequences from several sources including GenBank,reference sequence and etc., of the entire genome and then comparing it to a reference genome in order to detect the genetic variation such as single nucleotide variants (SNV), insertions, deletions,

inversions, rearrangements of genes, and copy number variant. To know about the content of the nucleotide in the DNA or RNA sequence.

At first, we have to input the nucleotide sequence, and identify whether it is DNA or RNA sequence and we count the number of nucleotide present in the nucleotides using the specific code and calculate the GC content of the sequence. If both the U (Uracine) and T (Thymine) is present in the sequence, then it is not a nucleotide sequence. If it has only U, then it is a RNA sequence else it is a DNA sequence. Then, count the number of ATGC (Adenine, Thymine, Guanine, and Cytosine) in a sequence.

Calculation of GC value

It is necessary to calculate the GC content of the DNA sequences in order to find the DNA repair in the normal nucleotide sequence that is caused by Mutation. To calculate the GC content,

$$\mathbf{Gc\_count = ((g + c) / (a + t + g + c)),}$$

Whereas g, c, a, t represents the number of guanine, cytosine, adenine and thymine nucleotides present in the mutated DNA strand. GC value is calculated from frequency of the occurrence of required gene in the DNA sequence.

Formation of phylogenetic tree

Usually, the formation of protein gap, where it is used to estimate the divergence between two sequences, and its usually measured in quantity of evolutionary, thereby calculating the genetic distances between DNA (or mRNA) sequences and to build a phylogenetic tree based on the distance matrix, which is probably used to contribute the level of effectiveness of the affected genome cells for the specific patient. It differs from every individual. The freely available datasets on genome cells and their specific heart disease sequence can be downloaded from the internet (eg. NCBI – National Center for Biotechnology Information). Then, we can make multiple alignment using the sequence of the various heart disease that are treated on the Clustal. Therefore, we can easily form the protein tree in the DNA sequencing of the mutated genes.

MODULE IMPLEMENTATION

Datasetsdownloaded in FASTA file format

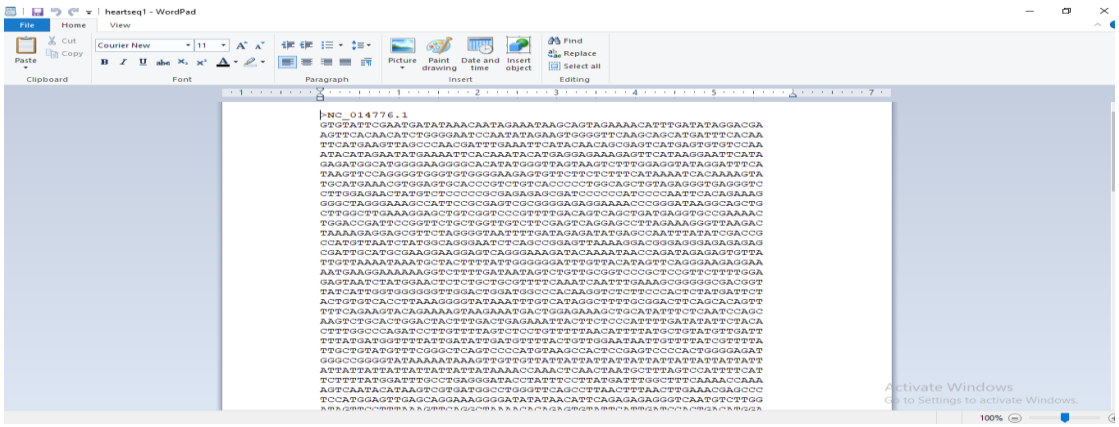


Figure.6: Downloading DNA Sequences in FASTA file format

Reading Fasta file for the specific DNAsequence

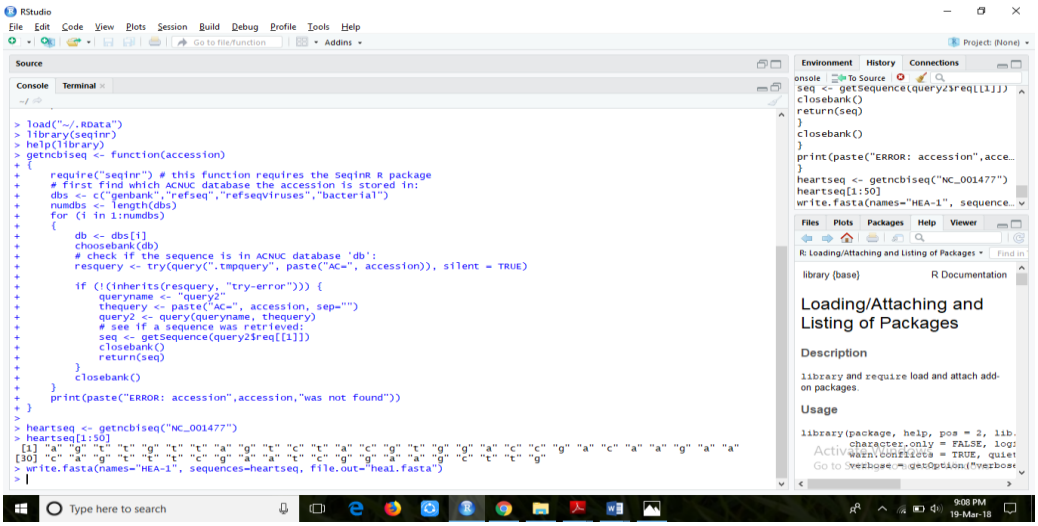


Figure.7: Reading FASTA file in R tool

Counting nucleotides for the datasets imported

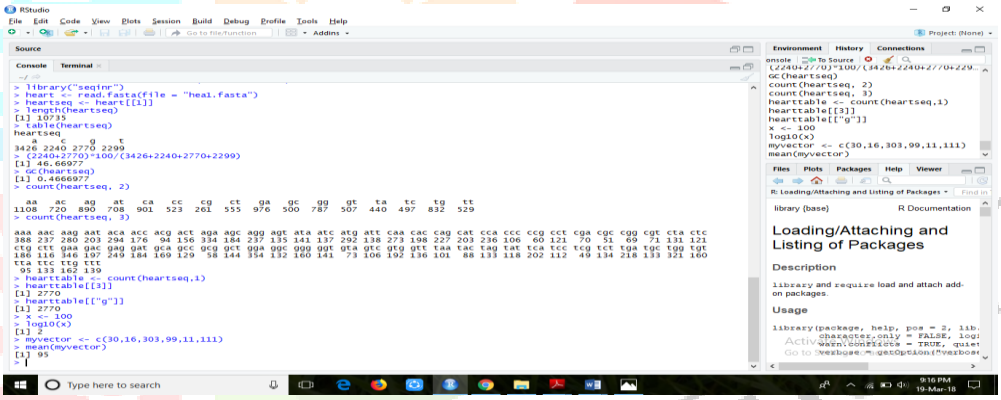


Figure.8: Counting of Nucleotides

Calculating GC value for imported datasets

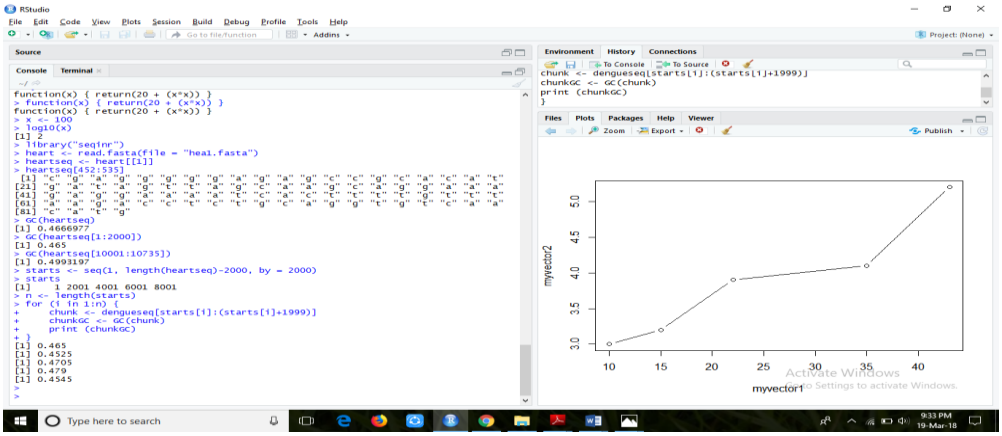


Figure.9: Calculation of GC value

Reading text base file and Plotting of graph

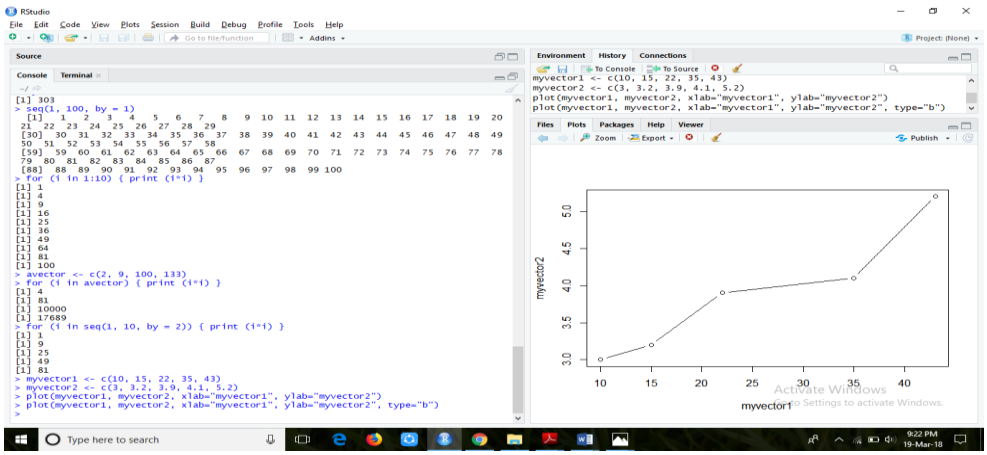


Figure.10: Plotting of graph

Graphical Representation of GC content versus Nucleotide start position

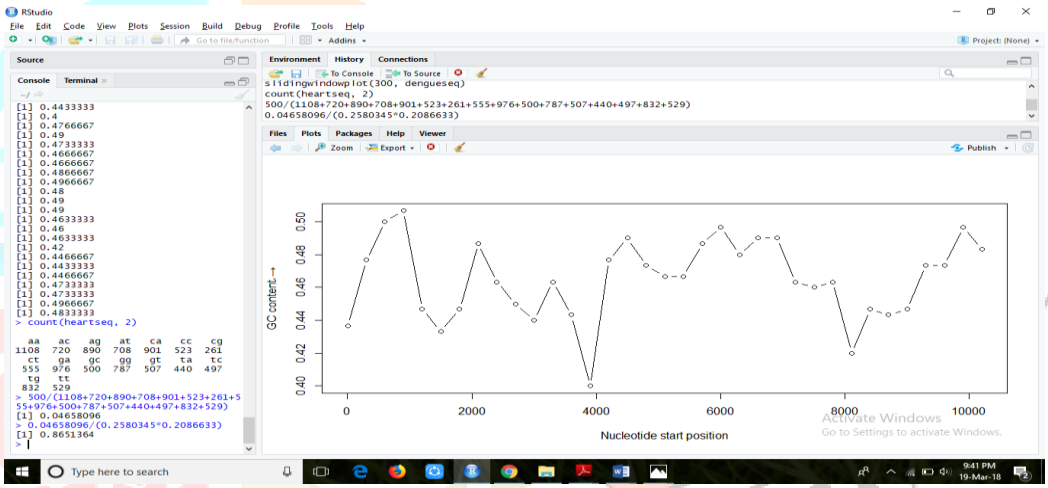
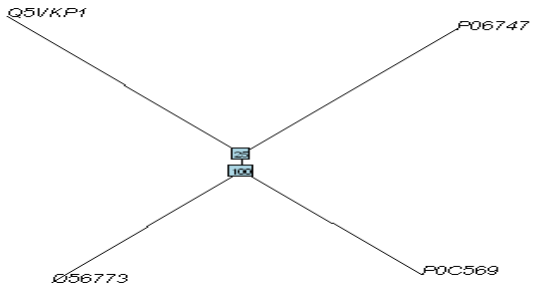


Figure.11: GC content versus Nucleotide start position

Formation of phylogenetic tree

Unrooted phylogenetic tree for protein sequences



Rooted phylogenetic tree for protein sequences

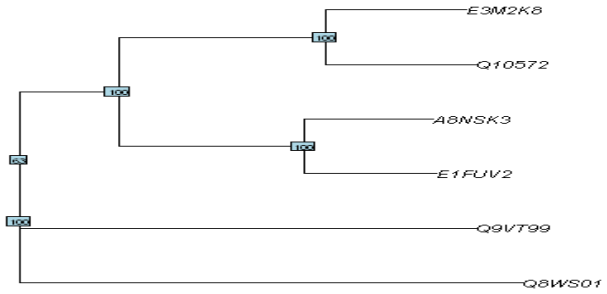


Figure.12: Formation of rooted and unrooted protein tree

Building a phylogenetic tree for DNA or mRNA sequences

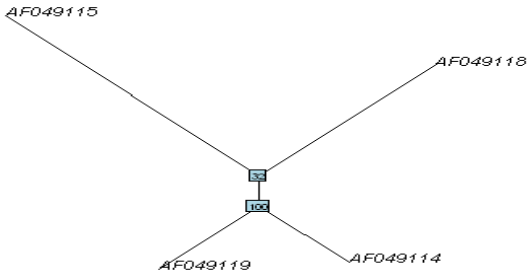


Figure.13: Formation of Phylogenetic tree

CONCLUSION

The study of DNA sequencing on gene cells and suggestion of affected genes for the patients is a huge and complex process since it also comprises best out of bioinformatics and big data analytics methodologies. And so the range of variation caused in the DNA nucleotide sequence is analyzed by calculating GC value of the Genome structure. Therefore, by this analysis, we can easily find out the defects in the DNA sequencing of the mutated genes. And also we can contribute the efficient phylogenetic tree for such kinds of heart

ACKNOWLEDGEMENT

References

1. Ismail Sahin Gul, Jens Staal, Paco Hulpiau, Evi De Keuckelaere, Kai Kamm, Tom Deroo, Ellen Sanders, Katrien Staes, Yasmine Driege, Yvan Saeys, Rudi Beyaert, Ulrich Technau, Bernd Schierwater and Frans van Roy “GENOME BIOLOGY AND EVOLUTION (GBE)” GC content of early metazoan genes and its impact on geneexpression levels in mammalian cell lines, February 2018.

2. Fernando Alvarez-Valin, Guillermo Lamolle, Giorgio Bernardi, Seccion Biomatematica, Facultad de Ciencias, Montevideo, Uruguay Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, I-80121 Naples, Italy “AN INTERNATIONAL JOURNAL ON GENES AND GENOMES” Isochores, GC3 and mutation biases in the human genome, June 2002.

3. Elhaik E, Landan G, Graur D. 2009. Can GC content at third-codon positions be used as a proxy for isochore composition? Mol Biol Evol. 2009**,** September 2010.