JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

Prediction of Human Genetic Disease based on Guanine - Cytosine Count

Annwesha Banerjee[1*], Anindya Sundar De[2], Rashbihari Halder[3], Gopal Basak[4], Agnish Majumder[5]

[1,2,3,4,5]Dept. of Information Technology, JIS College of Engineering, India

[*]Corresponding Author: annwesha.banerjee@jiscollege.ac.in

DOI: https://doi.org/10.26438/ijcse/v7i2.528531 | Available online at: www.ijcseonline.org

Accepted: 20/Feb/2019, Published: 28/Feb/2019

Abstract— Through the proposed method GC content of human DNA sequence have been calculated. The GC content plays a major role in disease prediction. Normally in a human genome the GC content is 35% to 60% , if found less than 35% then it indicates about some deficiency diseases like essential amino acid deficiency disease (mainly Alanine, proline, glycine); and if this content is found more than 60%, then it can be indication of some chromosomal or genetic diseases. So, based on the report of GC content a human can take some precautions to eradicate the probability of happening these kind of diseases.

***Keywords*—** Alanine, Cytosine, DNA, Guinine, Glycine, Proline

I. Introduction

Bioinformatics is the use of computers for the acquisition, management, and analysis of biological information. Bioinformatics is the interaction between computation and biology where computation is being used to biological data analysis and at the same time machine learning is one of the basic requirements for biological data computation [1]. Bioinformatics is emerging and advance branch of biological science, contain Biology mathematics and Computer Science. Genetic information is very valuable for different disease prediction and family risk analysis. The central dogma of biology holds that DNA from alleles at a genetic locus translates into proteins. DNA is treated as "blue print of life". It contains all the information to create life. DNA contains the information needed to create the amino acids sequences of proteins. The unit of building block of DNA Adenine (A), Cytosine (C), Guanine (G), and Thymine (T) are the four bases in DNA. A pairs with T that is 2H bond and C pairs with G that is 3H bond. In recent decades, exome sequencing has primarily been used in patient studies. The process of identification of genomic DNA regions encoding proteins is defined as gene prediction or gene finding.

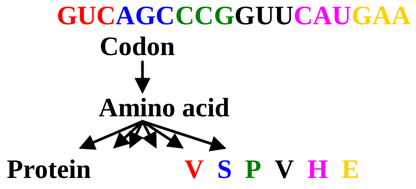


Figure 1: mRNA to Protein.

Protein is a linear sequence of amino acids, shown in the Figure 1 form a very long chain via peptide linkage. Gene is a segment of DNA. Inheritance pattern are the predictable pattern seen in the transmission of genes from one generation to the next and their expression in the organism that possesses them.

There is plenty of work in this field to predict disease by analyzing the gene sequences. Controlled by multiple, sometimes numerous, genes, the heredity diseases are genetically complex [2]. Even most

phenotypes are not monogenic.[3]. Health genome is a very successful project under the National Institute of Health and Department of Energy to predict cardiovascular diseases. [4]. Genomic tests which is very useful that can be performed without detectable risk or significant stress to the patients [5, 6].
Gene analysis is not only useful for disease prediction but also can be applied for preliminary care of the patient.[7, 8,9]. For prediction of cardiovascular disease assay thousands of genes simultaneously using micro array is also helps a lot.[11,12]. Clinical staging, gene expression profiling of the tumor can be to predict long-term disease recurrence and survival as well as possibly for planning treatment regimens[13,14,15]. The detection of atherosclerosis can be possible with the blood gene prediction which has been found in recent study[16]. The feasibility of screening for monogenic diseases across the genome within 50 h in a neonatal clinical setting has been proposed by Saunders et al. [17].

II. PROPOSED METHODOLOGY

Through our proposed method the count of GC content has been measured that will help for disease prediction. The pictorial diagram of the method has been depicted in figure 2.

The proposed method is actually composed of following five steps:

Step1: Collection of Human Genome data on which the computation has been experimented.

Sample dataset is given as:
Authority International Nucleotide Sequence
Database
Collaboration
Contact NCBI
Scope / transl_table qualifier
URL:
http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.
cgi?mode=c
Genetic Code [1]
Standard Code (transl_table=1)
AAs =
FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIM
TTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = ---M---------------M---------------M---------------

Base1 =
TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCC
AAAAAAAAAAAAAAAAGGGGGGGGGGGGGGG
G
Base2 =
TTTTCCCCAAAAGGGGTTTTCCCCAAAAGG
GGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGG
G
Base3 =
TCAGTCAGTCAGTCAGTCAGTCAGTCAGTC
AGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCA

Step 2: Storing the DNA sequences in file.

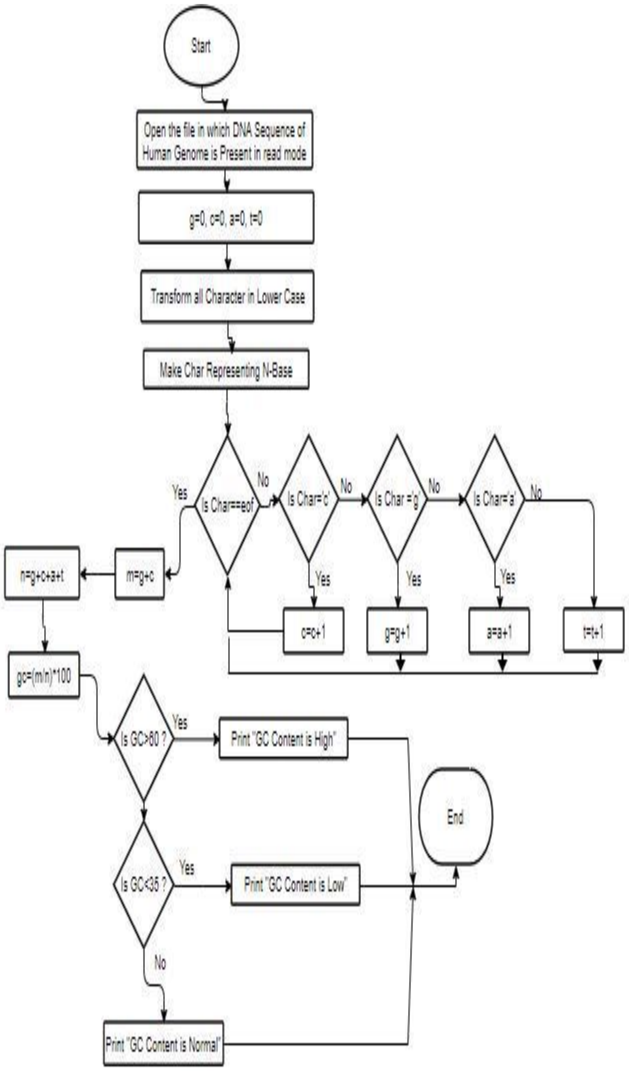
Step 3: Opening the file where human genome sequence is present

Step 4: Calculating the business logic

Step 5: Displaying the predictions.

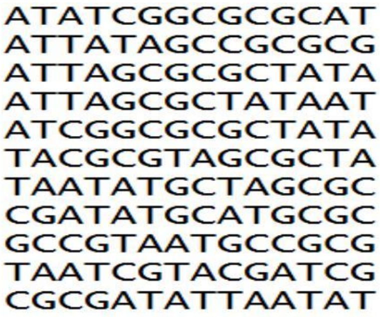
ALGORITHMIC APPROACH

```
Step 1:- Open the file in which the DNA sequence of a human genome is present, in read mode and store the data in a variable named gene.

Step 2:- [Initialize] g= 0, c= 0, a= 0, t= 0

Step 3:- Transform all the characters, present in the sequence, in lower case

Step 4:- Make 'char' representing each N-base present in the sequence

Step 5:- Repeat step 6 while end of file is not reached
Step 6:- If char= g then
                Set g= g+1
            If char= c then
                Set c= c+1
            If char= a then
                Set a= a+1
            If char= t then
                Set t= t+1
Step 7:- Set m= g+ c and n= g+ c+ a+ t
Step 8:- Calculate, gc= (m/n) * 100
Step 9:- If gc> 60
            Print "GC content is High"
         Else if gc< 35
            Print "GC content is Low"
         Else
            Print "GC content is Normal"
Step 10:- Exit
```

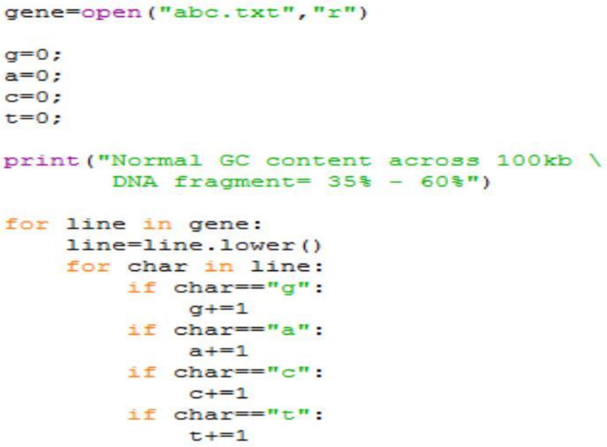

Start
Open the file in which DNA Sequence of Human Genome is Present in read mode
g=0, c=0, a=0, t=0
Transform all Character in Lower Case
Make Char Representing N-Base
Is Char==eof
Yes
No
Is Char='c'
No
Is Char ='g'
No
Is Char='a'
No
Yes
Yes
Yes
n=g+c+a+t
m=g+c
c=c+1
g=g+1
a=a+1
t=t+1
gc=(m/n)*100
Is GC>60 ?
Yes
Print "GC Content is High"
End
Is GC<35 ?
Yes
Print "GC Content is Low"
No
Print "GC Content is Normal"

RESULT ANALYSIS

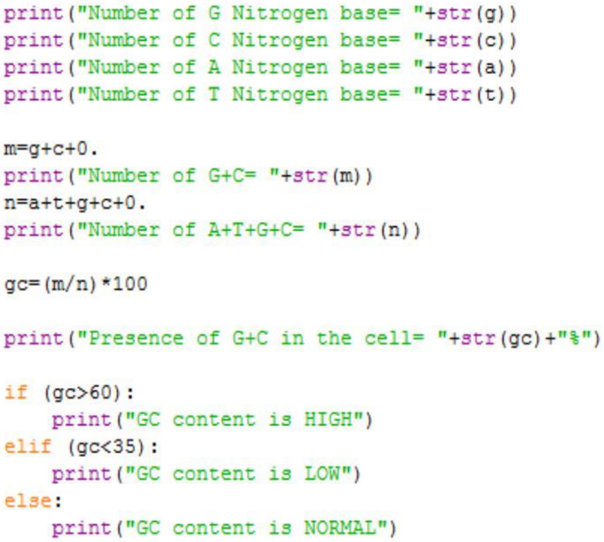
At first the file in which the Human DNA sequence is present, is opened in Read mode



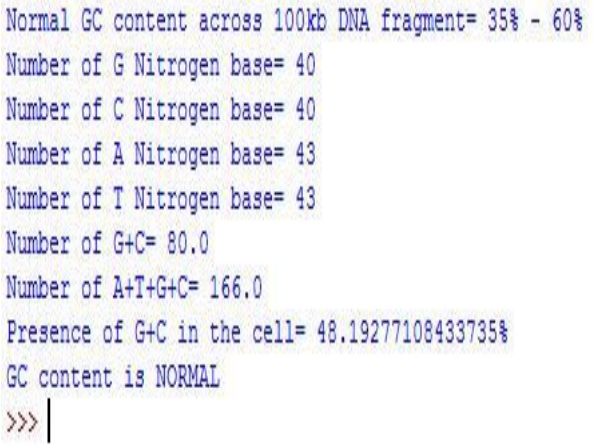
```
ATATCGGCGCGCAT
ATTATAGCCGCGCG
ATTAGCGCGCTATA
ATTAGCGCTATAAT
ATCGGCGCGCTATA
TACGCGTAGCGCTA
TAATATGCTAGCGC
CGATATGCATGCGC
GCCGTAATGCCGCG
TAATCGTACGATCG
CGCGATATTAATAT
```



```
gene=open("abc.txt","r")

g=0;
a=0;
c=0;
t=0;

print("Normal GC content across 100kb \
       DNA fragment= 35% - 60%")

for line in gene:
    line=line.lower()
    for char in line:
        if char=="g":
            g+=1
        if char=="a":
            a+=1
        if char=="c":
            c+=1
        if char=="t":
            t+=1
```

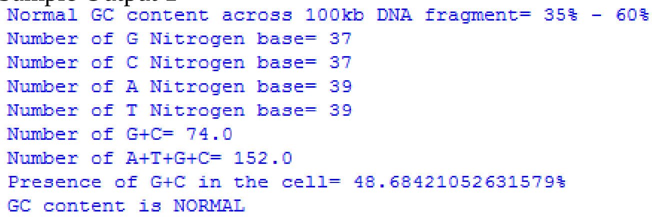


```
print("Number of G Nitrogen base= "+str(g))
print("Number of C Nitrogen base= "+str(c))
print("Number of A Nitrogen base= "+str(a))
print("Number of T Nitrogen base= "+str(t))

m=g+c+0.
print("Number of G+C= "+str(m))
n=a+t+g+c+0.
print("Number of A+T+G+C= "+str(n))

gc=(m/n)*100

print("Presence of G+C in the cell= "+str(gc)+"%")

if (gc>60):
    print("GC content is HIGH")
elif (gc<35):
    print("GC content is LOW")
else:
    print("GC content is NORMAL")
```



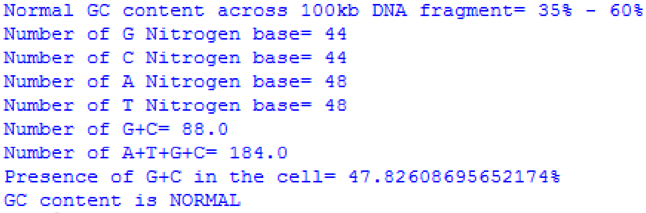
```
Normal GC content across 100kb DNA fragment= 35% - 60%
Number of G Nitrogen base= 40
Number of C Nitrogen base= 40
Number of A Nitrogen base= 43
Number of T Nitrogen base= 43
Number of G+C= 80.0
Number of A+T+G+C= 166.0
Presence of G+C in the cell= 48.19277108433735%
GC content is NORMAL
>>>
```

Sample Output 1



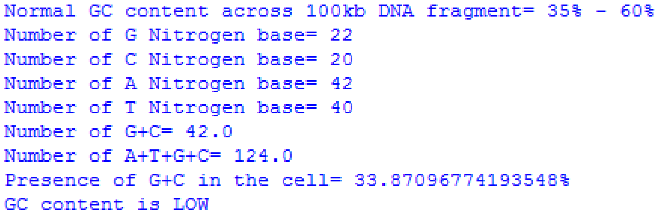
```
Normal GC content across 100kb DNA fragment= 35% - 60%
Number of G Nitrogen base= 37
Number of C Nitrogen base= 37
Number of A Nitrogen base= 39
Number of T Nitrogen base= 39
Number of G+C= 74.0
Number of A+T+G+C= 152.0
Presence of G+C in the cell= 48.68421052631579%
GC content is NORMAL
```

Sample Output 2



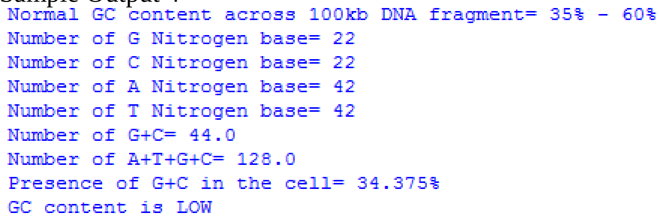
```
Normal GC content across 100kb DNA fragment= 35% - 60%
Number of G Nitrogen base= 44
Number of C Nitrogen base= 44
Number of A Nitrogen base= 48
Number of T Nitrogen base= 48
Number of G+C= 88.0
Number of A+T+G+C= 184.0
Presence of G+C in the cell= 47.82608695652174%
GC content is NORMAL
```

Sample Output 3



```
Normal GC content across 100kb DNA fragment= 35% - 60%
Number of G Nitrogen base= 22
Number of C Nitrogen base= 20
Number of A Nitrogen base= 42
Number of T Nitrogen base= 40
Number of G+C= 42.0
Number of A+T+G+C= 124.0
Presence of G+C in the cell= 33.87096774193548%
GC content is LOW
```

Sample Output 4



```
Normal GC content across 100kb DNA fragment= 35% - 60%
Number of G Nitrogen base= 22
Number of C Nitrogen base= 22
Number of A Nitrogen base= 42
Number of T Nitrogen base= 42
Number of G+C= 44.0
Number of A+T+G+C= 128.0
Presence of G+C in the cell= 34.375%
GC content is LOW
```

Sample run result 1, and 3 has been shown that the GC content is nearly 48.19%, 47.68 % and 48.87 % respectively.

So, GC content is normal as it is within 35-60%. In case of sample output 4 and 5 the GC count are 33.87% and 34.37% respectively which implies the GC count as low.

CONCLUSION

Bioinformatics is an emerging field of research mainly in the field of disease prediction and personalized drug inventions. Gene sequence analysis is a very effective process for disease prediction. Through the method a simplified prediction regarding the chances of disease based on the GC count has been proposed.

REFERENCE

[1] Baldi P and Brunak S (1998) Bioinformatics: The Machine Learning Approach, MITPress, Cambridge,MA.

[2] lazier AM, Nadeau JH & Aitman TJ (2002) Finding genes that underlie complex traits. Science 298 , 2345– 2349.

[3] Botstein D & Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for Men- delian disease, future approaches for complex disease. Nat Genet 33 , 228–237.

[4] Spencer G. International Consortium Completes Human Genome Project. Bethesda, MD: National Institutes of Health, 2003.

[5] Collins FS, Green ED, Guttmacher AE, Guyer MS, U.S. National Human Genome Research Institute. A vision for the future of genomics research. Nature 2003;422:835– 47.

[6] Bell J. Predicting disease using genomics. Nature 2004;429:453– 6.

[7] 7. Roses AD. Pharmacogenetics and the practice of medicine. Nature 2000;405:857.

[8] Williams RSaG.-C PJ. The genetics of cardiovascular disease: from genotype to phenotype. Dialogues in Cardiovascular Medicine 2004; 9:3–19.

[9] Guttmacher AE, Collins FS. Genomic medicine—a primer. N Engl J Med 2002;347:1512–20.

[10] Cook SA, Rosenzweig A. DNA microarrays: implications for cardiovascular medicine. Circ Res 2002;91:559 – 64.

[11] Goldsmith ZG, Dhanasekaran N. The microrevolution: applications and impacts of microarray technology on molecular biology and medicine (review). Int J Mol Med 2004;13:483–95.

[12] Napoli C, Lerman LO, Sica V, Lerman A, Tajana G, de Nigris F. Microarray analysis: a novel research tool for cardiovascular scientists and physicians. Heart 2003;89:597– 604

[13] van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999 –2009.

[14] Bild AH, Yao G, Chang JT, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 2006;439: 353–7.

[15]. Berchuck A, Iversen ES, Lancaster JM, et al. Patterns of gene expression that characterize long-term survival in advanced stage serous ovarian cancers. Clin Cancer Res 2005;11:3686 –96.

[16] Patino WD, Mian OY, Kang JG, et al. Circulating transcriptome reveals markers of atherosclerosis. Proc Natl Acad SciUSA 2005;102: 3423– 8.

[17] Saunders, M., Lewis, P. and Thornhill, A. (2012) Research Methods for Business Students. Pearson Education Ltd., Harlow.