JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

GC content Analysis using Machine Learning Algorithms

Dr. F. Amul Mary
Department of Computer Science
JMJ College for Women (A), Tenali

Abstract:

The genome in human body programs the blueprint of one's life. The genome sequence in human body provides the fundamental rules for human biology. Science makes every effort to reveal the laws of nature and critical understanding of the biology. Scientists in the life-science field and technological advancements helps the scientists to quickly create, store and analyze the data as fast as possible and as efficient as possible. The NCBI and other organizations maintain genome sequences, proteins, RNA, DNA and other information of all species as well as their behavioral data. There is a lot of data and translating these data into useful insights is the major concern. This is possible with big data technology that handles unstructured and semi structured and structured data. Big data plays a vital role in the field of Bioinformatics to extract meaningful information from large biological datasets to identify clinically actionable genetic variants for individualized diagnosis. Big data analytics helps the practitioners to give medical care to the patients from depiction to prediction with correct decision making capability. This paper aims at exploring the intersection between big data analytics and genomics particularly GC Content analysis which is useful for prediction and genome annotation. GC Content determination is useful in DNA and it affects the stability of DNA and secondary structure of mRNA. GC content contributes to the evolution rate of amino acid.

Keywords:

Big data Analytics, Genome, GC Content, NCBI

I. Introduction:

Big data is used to describe the collection of data which are huge in size like mountains. Big data is very large and complex. So, it cannot be processed using traditional data management tools efficiently. Big data is used to illustrate about voluminous amount of data in structured, semi structured and unstructured format that can mined to get relevant information, the data that cannot be processed using traditional approach [4]. The process of accessing and storing the large amount of data for analytics exists for quite a long time in the world. But, the perception about Big Data expanded in early 2000s by Dough Laney, the industry analyst about the mainstream of Big Data as 3 V's namely Volume, Velocity and Variety. This is illustrated in the figure.

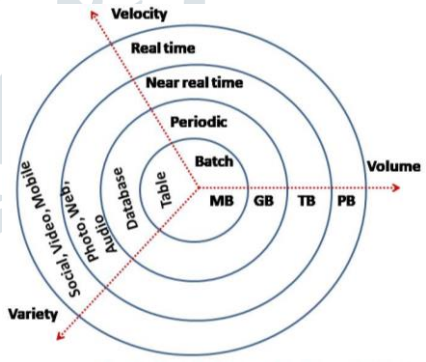


Fig1: 3V's of Big Data Characteristics

Apart from the above 3Vs, there are components added to big data like value, veracity, validity and variability [4].

Machine learning Algorithms learn the hidden patterns from the data predict the output and improve the performance from the experiences on their own and can make decisions.

R language has the best tools and library packages to work with machine learning algorithms. BGData suite of R was developed for the scientists to analyze extremely large genomic datasets with R environment. R

programme with its statistical heritage has plotting features and packages is one of the best language to analyze genomic data. Bioconductor and CRAN has array of specialized tools for performing genomics specific analysis.

II. Literature Review:

Ju Han Kim in the book entitled, ***"Genome Data Analysis"*** says that Genome sequence analysis unveils the genetic information of life which is the most classic methodology of the field. The data needed to analyze the genetic information is made to be available to everyone through TCGA project. This has to be given access for every researcher in the field of genomics [1].

Ward Jonathan and Barker defined the big data in terms of volume, complexity, and technology. Their definition is "*Big data is the technology involves storing and analyzing huge amounts of data collected from a variety of sources*". Machine learning methods are used for analyzing data [2].

Karen.Y.He, Dongliang.Ge, Max.M.He suggested different ways to manipulate manage and analyze genomic and clinical data. This will help in administering accurate medicines. The authors introduced big data tools that could assist in identifying the genetic variations [3].

Petr Smarda, Lucie Horova, Ettore Pacini, et.al reported through their analysis that GC content is predicted to affect genome functioning significantly. Analysis on the genomic GC content of 239 species was performed. The samples consisted of 70 to 78 families belonging to monocots. The variation in GC content of the monocots is found to show a variation between 33.6% and 48.9%. [5].

Steve G P, Buntrock J D said that the availability of Electronic Health Record Systems helps anyone residing on the globe to have a glimpse of the variety of clinical data that is available in abundance. This expedites the researchers to carry out their research in a more precise manner [10].

Dennis A. Benson, David L. Wheeler, et.al says about the availability of a database by the name GenBank. This database provides information about the nucleotide sequences that are available for the public to use. This database was created with the information gathered from various laboratories and other projects. One can gain access to this database through the **NCBI** (National Center for Biotechnology Information) Portal [13].

Luciano Brocchieri stated that the GC content that is present at the three codon positions **(GC1, GC2, and GC3)** shows linear variations with the other GC contents of the genes. The authors also state that the variation observed in the first two codon positions is comparatively less than the variation observed at the third codon position [14]. The GC content observed values lies between as GC3=0.0 and GC3=1.0.

II. Data Analysis Process: Data analysis is a process of inspecting, cleansing, transforming, and modeling the data with the aim to discover useful information and draw the conclusion and supporting decision making. Data analysis is a process for obtaining raw data and converting it into information useful for decision making by users. Data is collected and analyzed to answer queries, test hypothesis or disapprove theories [8].

The process of Data Analysis includes: Data Requirements, Data Collection, Data Processing, Data Cleaning, Exploratory Data Analysis, Modeling and Algorithms, Data Product and Communication of Visualized Report. The phases are iterative in nature. The feedback of later phases may result in earlier phases [8]. The phases are shown in the figure.

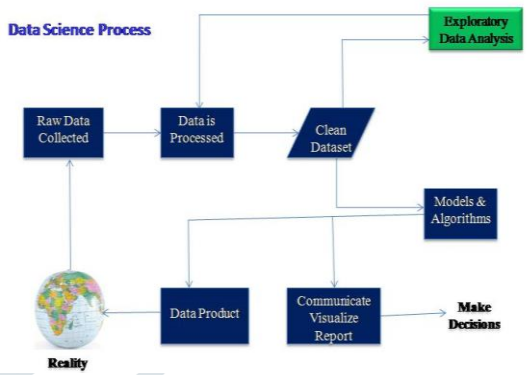


Fig2: Process of Data Analysis - Flowchart

IV. Bioinformatics in big data analytics: Bioinformatics have received considerable attention for the past few years, wherein computationally efficient methods have been developed for genomic data analysis like location of coding regions in a genome sequence. Big data infrastructure exhibits challenges and provided feasible opportunities to develop an effective and valuable approach to identify clinically actionable genetic variants for individualized diagnosis. Due to the effectiveness of big data, it has been widely used in various research fields [12].

Big data refers to new technological tools that deliver potential capabilities for managing and processing massive and varied datasets. On a population level, big data provides the possibility to conduct large scale

exploration of clinical consequences to uncover hidden patterns and correlations [10]. Big data analytics helps the practitioners to give medical care to the patients from depiction to prediction with correct decision making capability. In clinical practice, diseases description is collected from different streams such as pathology, images from scanning report, electrophysiology and genomics to give care for the patients.

IV.I) The central dogma of Genomics: The field of study about genome is called "*Genomics"*. The term "*genomics"* was proposed by Tom Roderick in 1986. But, Rosalind Franklin confirmed helical structure to DNA and James D. Watson and Francis Crick officially published the concept of helical structure of DNA way back in 1953. Marshall Nirenberg and Philip Leder discovered the triplet nature of genetic code and they worked with 54 out of 64 triplet codons in their experiments. Frederick Sanger and his colleagues were the first to sequence DNAs. The sequenced genome data had to be processed to extract genetic information from genes [12].

The genome includes both the coding regions and non-coding regions DNA. The coding region of a gene also known as CDS meaning coding Sequence which is the portion of a gene's DNA or RNA that codes for protein. The studying of length, composition, regulation, splicing and structures of coding regions compared to non-coding regions over different species and time periods provide information regarding gene organization and evolution of prokaryotes and eukaryotes. This again helps in mapping the human genome and developing gene therapy [11].

DNA molecules are based on four nucleotides A, T, G and C. The DNA sequence contains the instructions to build up protein and other molecules that the cell needs to carry the daily work. Instructions in DNA first translate into RNA and RNA is translated into Proteins.

IV.II) Genome Sequence Formats: Understanding the structure of different data formats are one of the first requirements in genome sequence data analysis. There are different formats like plain sequence, fastq, embl, fasta, GCG and Genbank format. The data was collected from NCBI website for Homo sapiens membrane metalloendopeptidase transcript of different variants, mRNA in fasta format.

Plain Sequence Format: Sequences in plain format contains International Union of Pure and Applied Chemistry characters and spaces but not numbers [12].

An example sequence in plain format is:

```
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA
```

Fig3: Plain sequence format

Genbank Format: The Sequence file in this format contains many sequences [12]. It starts with the word "LOCUS" and a number of annotation lines. The starting of the sequence is marked by "ORIGIN" and ending of the sequence is marked by "//"

An example sequence in GenBank format is:

```
LOCUS       AB000263               368 bp    mRNA    linear   PRI 05-FEB-1999
DEFINITION  Homo sapiens mRNA for prepro cortistatin like peptide, complete
            cds.
ACCESSION   AB000263
ORIGIN
        1 acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg
       61 ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg
      121 caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc
      181 aggccagtgc cgggccctc ataggagagg aagctcggga ggtggccagg cggcaggaag
      241 gcgcacccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga
      301 agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca
      361 gacctgaa
//
```

format contains many sequences [12]. It starts with the word "LOCUS" and a number of annotation lines. The starting of the sequence is marked by "ORIGIN" and ending of the sequence is marked by "//"

An example sequence in GCG format is:

```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
AB000263  Length: 368  Check: 4514  ..
      1  acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg
     61  ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg
    121  caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc
    181  aggccagtgc cgggccctc atagggagagg aagctcggga ggtggccagg cggcaggaag
    241  gcgcacccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga
    301  agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca
    361  gacctgaa
```

Fig 4: GenBank Format

GCG Format: Sequence file in GCG format contains exactly one sequence. It begins with annotation lines. The starting of sequence is marked by ".." characters and it contains sequence identifier, sequence length and a checksum.

```
>NM_001354644.1 Homo sapiens membrane metalloendopeptidase (MME), transcript variant 5, mRNA
AGGAGAGCGAGCTGAGGGAGAAAGGTCCAAAGGGCGCGAGCGCCCAGGGCGCTGGGAGCCCGTGGGACGG
GCACGGGGAGGGCAGAGCCAGCCGAGGGGAGCGGCTGCCGGGAGGTGCGCTGAGCAGGGTCCGCAGCTAA
GGTCCAGCGCGATCCGAGCGCCCAGGACCCTGCGGCCGTGGAGGAGGCGTGCCCTGGGAAGGAGCCGCTA
CTGGGACCTGAAGAGTGAGGGGAGGGAGAGGGGCCCAGAGGGTGACCTGGAGGAGGGCTCTGGAAGTCAC
GTCAGGTTGGCTCTTCAGGTTCATTTCCATAGTTCCCTGCGGCCTCTGCCTTGGGGAGTTATGTTTTGTT
ACCGAGATCCGCGCTACCAGATTGCACCGGGGCTGATTTGGGGGCTGGGAATTTGCCATTCTGCTGTACA
GACACTGATTTTTTTTTCTTCTTTTTAAAAAGCAAGATTTTAGGTGATGGGCAAGTCAGAAAGTCAGATG
GATATAACTGATATCAACACTCCAAAGCCAAAGAAGAAACAGCGATGGACTCCACTGGAGATCAGCCTCT
CGGTCCTTGTCCTGCTCCTCACCATCATAGCTGTGACAATGATCGCACTCTATGCAACCTACGATGATGG
```

Fig 5: GCG Format

Fasta Format: The FASTA format contains several sequences. The sequences in this format begins with a single line that contains description and greater than symbol ">" followed by sequence data [12].

***IV. III).* GC content analysis of genome sequence:** One of the most fundamental properties of a genome sequence is its GC Content. GC Content analysis is the percentage of Nucleotides in genome sequence that is Guanines and Cytosine in a DNA and RNA molecule. GC can be calculated for fragment of DNA or RNA or even for entire genome. GC content is an important factor in shaping amino acid compositions. It is possible to identify the factors which cause the amino acid usage of proteins during the development process of human being, if the amino acid composition is known [15].

GC content is useful for prediction and genome annotation. GC Content determination is useful in DNA. GC content affects the stability of DNA and secondary structure of mRNA. GC contents contribute to the evolution rate of amino acid. The causes for variation in GC content is that GC contents are correlated with many genomic features like replication timing [16], aerobiosis, that is aerobic prokaryotes display a significant increment in genomic GC in relation to anaerobic ones and there is link between metabolic character and GC content [17].

The calculation of GC Content for the Genomes as follows: The GC content of the gene is defined as the frequency of nucleotides that are guanine or cytosine. The G + Content at the 1st position of synonyms codons are the fraction of codons that are synonyms at the 1st codon position, which have either guanine or cytosine at the first position. Similarly, the G + Content at the 2nd position of synonyms codons are the fraction of codons that are synonyms at the 2nd codon position, which have either guanine or cytosine at the second position and the G + Content at the 3rd position of synonyms codons are the fraction of codons that are synonyms at the 3rd codon position, which have either guanine or cytosine at the third position. Here, in this paper, GC Content is calculated for 8 selected genomes of Homo sapiens membrane metalloendopeptidase transcript of different variants, mRNA in fasta format is calculated using R Programming in RStudio environment.

Table1: Calculation of A, T, G, C and Total length of Nucleotide

SNo	Genome ID	Total Adenine	Total Thymine	Total Guanine	Total Cytosine	Total Nucleotide
1	NM_001354644.1	390	291	357	305	1343
2	NM_001354642.1	1811	1084	1127	1712	5734
3	NM_001354643.1	1807	1105	1174	1671	5757
4	NM_007288.3	1820	1115	1205	1678	5818
5	NM_007287.3	1780	1094	1113	1678	5665
6	XM_006713647.4	2318	1416	1454	2448	7636
7	XM_011512856.2	1462	831	857	1343	4493
8	XM_011512857.2	1447	827	855	1335	4464
9	NC_060928.1	25299	26891	17359	13968	83517
10	NC_060927.1	61130447	60661558	39576837	39737106	201105948

Table2: Calculation of GC1, GC2 and GC3

	SNo	Genome ID	GC1	GC2	GC3
1	1	NM_001354644.1	0.4825	0.5133	0.4675
2	2	NM_001354642.1	0.3856	0.4071	0.3584
3	3	NM_001354643.1	0.3959	0.3658	0.4033
4	4	NM_007288.3	0.3988	0.4198	0.3692
5	5	NM_007287.3	0.3896	0.4126	0.3601
6	6	XM_006713647.4	0.3759	0.3831	0.3923
7	7	XM_011512856.2	0.3757	0.3404	0.3767
8	8	XM_011512857.2	0.3768	0.4119	0.3387
9	9	NC_060928.1	0.375	0.3748	0.3748
10	10	NC_060927.1	0.3944	0.3943	0.3942

Fig7: Graphical representation of GC1

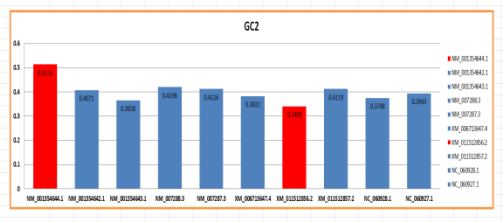


Fig8: Graphical representation of GC2

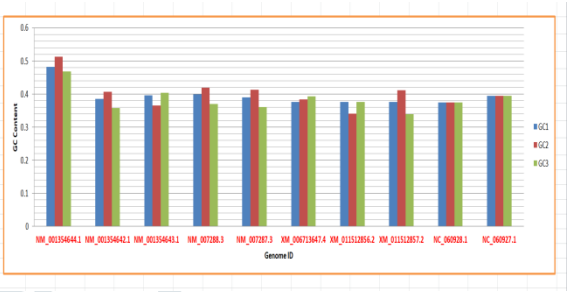
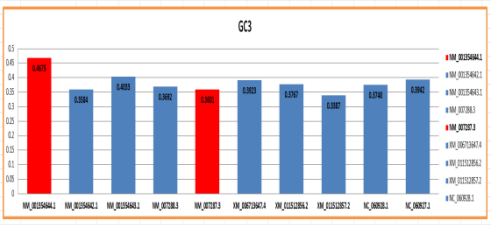
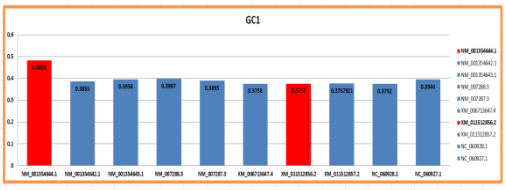


Fig9: Graphical representation of GC3





GC content is the proportion of DNA strand, RNA strand, gene region chromosome or genome that is Guanine (G) and Cytosine(C) rather than adenine (A) and Thymine (T). It is observed from the figure that GC content at the 1st codon position value ranges from 0.37 to 0.48 (that is 37% to 48%) and the GC content at the 2nd position value ranges from 0.34 to 0.51(that is 36% to 51%) and GC content at the 3rd position value ranges from 0.33 to 0.46 (that is 33% to 46%) for the selected genome sequences of Homo sapiens membrane metalloendopeptidase transcript of different variants, mRNA. The GC content of organisms is a highly variable trait.

V. References:

[1] Ju Han Kim, (2019), "Genome Data Analysis", learning materials in Biosciences book series (LMB), doi: 10.1007/978-981-13-1942-6_2.

[2] Ward Jonathan Stuart, Barker Adam, (2013), "Undefined By Data: A Survey of Big Data Definitions", 1309.5821

[3] Karen Y He, Dongliang Ge, Max M He, (2017), "Big Data Analytics for Genome Medicine", 18 (2): 412, doi: 10.3390/ijms18020412.

[4] F. Amul Mary, (2017), "A Survey of Big Data Analytics-It's Challenges" in International Journal of Advanced Research Trends in Engineering and Technology (IJARTET), ISSN:2394-3777(Print), ISSN:2304-3785(Online), Volume-4, Special Issue 21, doi:10.20247/IJARTET.2017.04080046

[5] Petr Smarda, Petr Bures, Lucie Horova, LLia J.Leitch, Ladislav Mucina, Ettore Pacini, Lubomir Tichy, (2014), "Ecological and Evolutionary significance of genomic GC Content diversity in monocots", https://doi.org/10.1073/pnas.1321152111

[6] F. Amul Mary, S. Jyothi, (2020), "Geometric Feature Extraction for Detecting Carcinoma in Three Dimensional MR Images through Machine Learning Algorithms" in International Conference on "Advances in Computational and Bio Engineering", ISSN: 2662-3447(P), 2662-3455(e), in the book "Learning and Analytics in Intelligent System 16", ISBN: 978-3-030-46942-9, ISBN: 978-3-030-46943-6(eBook), Volume-2, https://link.springer.com/chapter/10.1007/978-3-030-46943-6_45

[7] F. Amul Mary, S. Jyothi (2021), "Percentage Concentration of Nucleotides in Genome Data of SARS-Corona Viruses" in International Journal of Advanced Research in Engineering and Technology(IJARET), ISSN(P): 0976-6480; ISSN(O): 0976-6499; Volume 12, Issue 2, pp:411-421; Article ID: IJARET_12_02_039; Impact Factor: 10.9475; doi:10.34218/IJARET.12.2.2020.039

[8] F. Amul Mary, S. Jyothi, (2020), "Data analysis on Corona Virus Disease (Covid-19)- Its Challenges in the Midst of Pandemic", in "Scopus Indexed Journal – Solid State Technology", ISSN: 0038-111X; Vol.63, Cite Score: 0.3; Pg:2613– 2622. http://solidstatetechnology.us/index.php/JSST/article/view/2003

[9] F. Amul Mary, S. Jyothi (2020), "Percentage Nucleotide Concentration and Classification of SARS-Corona Viruses" in International Conference on "Advances in Computational and Bio Engineering" in the book, "Lecture Notes in Networks and Systems" https://link.springer.com/chapter/10.1007/978-981-16-1941-0_13

[10] Steve G Peters, James D Buntrock,(2014),"Big data and the electronic health record", J.Ambul Care Manag,37:206–210, https://pubmed.ncbi.nlm.nih.gov/24887521/

[11] Genome Size: https://en.wikipedia.org/wiki/Genome

[12] http://hdl.handle.net/10603/457967

[13] David L. Wheeler, Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Ron Edgar, Scott Federhen, (2007), "Nucleic Acid Research", Volume 35, Issue Suppl_1, Pages D5-D12.

[14] Luciano Brocchieri, (2014), "The GC content of the Bacterial Genomes", Phylogen Evolution Biol 2014, 2:1 DOI: 10.4172/2329-9002.1000e108

[15] Brooks D J, Fresco J R,(2002), "Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor", Mol Cell Proteomics 1:125. doi: 10.1074/mcp.M100001-MCP200.

[16] Deschavanne, P, Filipski, J, (1995), "Correlation of GC content with replication timing and repair mechanisms in weakly expressed E. coli genes", Nucleic Acids Res. 23, 1350–1353. doi: 10.1093/nar/ 23.8.1350.

[17] Naya, H., Romero, H., Zavala, A., Alvarez, B., and Musto, H, (2002), "Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes", J. Mol. Evol. 55, 260–264. doi: 10.1007/s00239-002-2323-3