HUMAN GENOME ANALYSIS

CS6003 – BIG DATA ANALYTICS

Submitted by

Sasmitha M S (2020103044)

Syed Junaid Ali B (2020103579)

in partial fulfilment of the requirements for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

COLLEGE OF ENGINEERING, GUINDY

ANNA UNIVERSITY: CHENNAI 600 025

MAY 2023

ANNA UNIVERSITY: CHENNAI 600 025

BONAFIDE CERTIFICATE

Certificate that this project request titled **“HUMAN GENOME ANALYSIS”** is the bonafide work of “**SASMITHA M S (2020103044), SYED JUNAID ALI B (2020103579)”,** who carried out the project work under my supervision, for the fulfilment of the requirements as part of the CS6003 – Big Data Analytics.

Dr. S. VALLI

PROFESSOR & HEAD

HEAD OF THE DEPARTMENT

Department of Computer Science
and Engineering
Anna University
Chennai – 600 025

Dr. AROCKIA XAVIER ANNIE

ASSISTANT PROFESSOR

SUPERVISOR

Department of Computer Science
and Engineering
Anna University
Chennai - 600025

ACKNOWLEDGEMENT

Foremost, we would like to express our sincere gratitude to our project guide, **Dr. Arockia Xavier Annie**, Assistant Professor, Department of Computer Science and Engineering, College of Engineering Guindy, Chennai for her constant source of inspiration. We thank her for the continuous support and guidance which was instrumental in taking the project to successful completion.

We are grateful to **Dr. S. Valli**, Professor and Head, Department of Computer Science and Engineering, College of Engineering Guindy, Chennai for her support and for providing necessary facilities to carry out for the project.

We would also like to thank our friends and family for their encouragement and continued support. I would also like to thank the Almighty for giving us the moral strength to accomplish our task.

SASMITHA M S **SYED JUNAID ALI B**

ABSTRACT

Genomic sequences encode the blueprint of life, providing valuable insights into the genetic makeup of organisms. Understanding the composition of proteins and nucleotides within a genomic sequence is crucial for unravelling their functional significance and evolutionary implications. This project aimed to analyze the protein and nucleotide composition in a given genomic sequence and draw meaningful conclusions about its genetic characteristics. The implementation involved parsing the genomic sequence and identifying protein and nucleotide sequences using predefined mappings. The frequency of occurrence of each protein sequence was computed, enabling the determination of their respective percentage compositions. Additionally, the distribution of nucleotides (A, G, T, C) and their percentages within the sequence were calculated. The (A+T)/(G+C) ratio, a measure of nucleotide bias, was also determined.

The results revealed the percentage composition of each protein, providing insights into the abundance of specific amino acids within the genomic sequence. The (A+T)/(G+C) ratio shed light on the nucleotide bias, potentially indicating variations in the genetic content and evolutionary patterns. Furthermore, the analysis of nucleotide composition highlighted the presence and relative abundance of A, G, T, and C within the sequence. The findings showcased the genetic diversity and organization of the genomic sequence under investigation. The visual representation of protein composition through a bar plot facilitated a comprehensive understanding of the relative abundances of different proteins. The descriptive statistics, including the average sequence length, longest sequence length, and shortest sequence length, provided additional insights into the structural characteristics of the genome.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

Advancements in genomic sequencing technologies have revolutionized our understanding of the genetic makeup of organisms. Genomic sequences contain a wealth of information encoded within the arrangement of nucleotides, forming the blueprint for the development, functioning, and evolution of life. Analyzing the composition of proteins and nucleotides within genomic sequences provides valuable insights into the molecular basis of biological processes, genetic diversity, and evolutionary relationships.

The primary focus of this project is to investigate and analyze the protein and nucleotide composition of a given genomic sequence. Proteins, composed of amino acids, are the workhorses of cellular processes, carrying out essential functions such as enzymatic catalysis, cell signaling, and structural support. Understanding the relative abundance and distribution of different proteins within a genomic sequence offers crucial insights into the biological processes at play and potential functional implications. Additionally, exploring the composition of nucleotides, the building blocks of DNA and RNA, provides valuable information about the genetic content and organization of the genomic sequence. Nucleotide composition analysis allows us to identify biases, such as the preference for certain nucleotides at specific locations, which can provide clues about evolutionary adaptations, regulatory elements, or functional constraints.

To accomplish this analysis, computational techniques are employed to handle the vast amounts of genomic data efficiently. Data manipulation and analysis methods are applied to extract meaningful patterns, calculate statistical measures, and visualize the results. By utilizing these techniques, we can gain insights into the distribution, diversity, and significance of proteins and nucleotides within the genomic sequence. The outcomes of this project provide a deeper understanding of the genetic landscape of the studied organism. By deciphering the protein composition, we can identify key proteins and their potential functional roles. Likewise, exploring nucleotide composition sheds light on the genetic variations and regulatory elements that influence gene expression and phenotype. These findings contribute to our knowledge of the organism's biology, evolution, and potential relationships with other species.

1.1. OVERALL OBJECTIVES

- Analyze the composition of proteins and nucleotides in a genomic sequence.
- Determine the percentage composition of proteins and nucleotides in the genomic sequence.
- Calculate the (A+T)/(G+C) ratio as a measure of genetic content.
- Investigate the genetic landscape and identify key proteins involved in biological processes.
- Explore the diversity and distribution of proteins and nucleotides in the genomic sequence.
- Gain insights into the genetic variations and regulatory elements present in the sequence.
- Understand the relationship between protein composition and functional roles in cellular processes.
- Apply computational techniques for data manipulation, statistical measures, and visualization.
- Implement code for efficient data processing and analysis of large genomic sequences.
- Provide meaningful insights into the genetic content and organization of the genomic sequence.
- Contribute to the understanding of fundamental processes in biology and evolution.
- Uncover the molecular components and arrangements that contribute to phenotype and genetic traits.

1.2. PRESENT ISSUES

i. Large Dataset Size: One of the challenges in this project is dealing with large genomic sequences, which can be several gigabytes or even terabytes in size. Handling such massive datasets requires efficient computational resources and optimized algorithms to ensure timely and accurate analysis.

ii. Data Processing Time: Analyzing and processing large genomic sequences can be time-consuming, especially when performing complex calculations and computations. The project needs to address techniques and optimizations to reduce the processing time and enhance efficiency.

iii. Memory Management: Working with large datasets requires careful memory management to avoid memory overflow or excessive memory usage. It is essential to design algorithms that can handle data in smaller chunks or employ techniques like lazy loading and parallel processing to optimize memory usage.

iv. Scalability: The project should consider scalability to accommodate future expansion and inclusion of more extensive genomic datasets. The ability to handle increasing dataset sizes and computational requirements is crucial for long-term viability and relevance of the project.

v. Data Accuracy and Quality: Genomic data can have inherent noise, errors, or missing information, which can impact the accuracy and reliability of the analysis. Ensuring data quality and implementing appropriate data cleansing techniques are important to obtain meaningful and accurate results.

vi. Interpretation of Results: Analyzing the composition of proteins and nucleotides in a genomic sequence can provide a vast amount of data. The project needs to address the challenge of interpreting and presenting the results in a meaningful way that can be easily understood by researchers and domain experts.

vii. Integration with Existing Tools and Databases: Integrating the project's findings and results with existing bioinformatics tools and databases is crucial for collaboration, cross-referencing, and further analysis. Ensuring compatibility and interoperability with established systems is important for maximizing the impact of the project.

viii. Privacy and Security: Genomic data contains sensitive and personal information. It is essential to implement robust security measures to protect the privacy and confidentiality of the data throughout the project's lifecycle, including data storage, analysis and dissemination.

ix. Ethical Considerations: Genomic research raises ethical considerations related to data usage, consent, and potential implications. The project should adhere to ethical guidelines and regulations to ensure responsible and ethical practices in handling genomic data.

x. Communication and Collaboration: Effective communication and collaboration with domain experts, biologists, and other stakeholders are critical for understanding the research objectives, incorporating domain-specific knowledge, and translating the findings into meaningful insights.

1.3. PROBLEM STATEMENT

Genomic analysis plays a crucial role in various fields such as genetics, bioinformatics, and medical research. With the advancements in sequencing technologies, the size and complexity of genomic datasets have grown exponentially. Analyzing and characterizing these large genomic sequences pose several challenges, including efficient data processing, memory management, data accuracy, result interpretation, integration with existing tools and databases, privacy and security concerns, ethical considerations, and effective communication with domain experts.

The problem at hand is to develop a comprehensive solution that addresses these challenges and enables efficient analysis and characterization of large genomic sequences. The solution should be able to handle massive datasets, optimize processing time, ensure accuracy and quality of results, and provide meaningful interpretations and visualizations. It should also seamlessly integrate with existing bioinformatics tools and databases, respecting privacy and security guidelines. Additionally, ethical considerations such as data anonymization and informed consent need to be taken into account.

The objective is to develop a robust framework that empowers researchers and scientists to gain valuable insights into the composition of proteins and nucleotides in genomic sequences. By addressing the present issues in genomic analysis, the solution will facilitate advancements in fields like personalized medicine, genetic research, and disease understanding. Ultimately, the project aims to contribute to the broader scientific community by providing a scalable, efficient, and user-friendly platform for genomic data analysis and interpretation.

1.4. ORGANIZATION OF THE THESIS

The thesis is organized into several sections to provide a structured presentation of the research work.

In the Introduction (Chapter 1), the overall objectives of the project are outlined, highlighting the main goals and aspirations. The present issues in the field are discussed, setting the context for the research. A comprehensive problem statement is presented, clearly defining the specific challenges addressed by the study. The organization of the thesis is also described, giving a brief overview of the subsequent chapters. Chapter 2 provides a summary of related work, presenting an overview of existing research and studies relevant to the project. A consolidation table is included, summarizing the key findings and insights from the reviewed literature.

The System Design (Chapter 3) elaborates on the design and architecture of the system. An architecture diagram is presented, illustrating the components and their relationships. The process of dataset collection, data cleaning, feature engineering, dataset splitting, feature selection, classification, and evaluation are described in detail, providing a comprehensive understanding of the system design. In Chapter 4, the results and discussions are presented. The platform and tools used for the project are introduced. A detailed description of the dataset is provided, including its characteristics and size. The data preprocessing steps are explained, highlighting the methods employed to clean and prepare the dataset. The implementation details and code are presented, allowing for a better understanding of the technical aspects. Finally, the obtained results are discussed, providing insights and interpretations of the findings.

The Result Analysis (Chapter 5) focuses on the evaluation parameters used to assess the results. It highlights the contributions made by the research, emphasizing the significance and impact of the findings. Chapter 6 concludes the thesis and presents future work. The conclusion section summarizes the key findings and contributions of the research. Future work is discussed, suggesting potential directions for further exploration and improvement.

The thesis is supported by a list of references, ensuring that proper credit is given to the relevant sources and prior studies. The organization of the thesis ensures a logical flow of information, enabling readers to follow the research process and understand the outcomes and implications of the study.

CHAPTER 2
RELATED WORKS

"Genomic analysis in the age of human genome sequencing" by Lappalainen et al. (2019):[1] This paper discusses the advancements in genomic analysis with the emergence of human genome sequencing. It highlights the importance of analyzing genomic data to understand genetic variations, gene expression, and their impact on human health. The study explores various genomic analysis techniques and their applications in studying complex diseases.

"Parallel human genome analysis: microarray-based expression monitoring of 1000 genes" by Schena et al. (1996):[2] The paper presents a microarray-based approach to monitor gene expression in human genomes. The study focuses on the simultaneous analysis of multiple genes and their expression patterns using microarrays. It demonstrates the potential of this technology for high-throughput genomic analysis and its application in understanding gene expression profiles.

"The human genome project" by Olson (1993):[3] This paper provides an overview of the Human Genome Project, a landmark initiative aimed at sequencing and mapping the entire human genome. It discusses the objectives, challenges, and potential benefits of the project in advancing genomic analysis. The study emphasizes the importance of comprehensive genomic data for understanding human biology and diseases.

"Initial sequencing and analysis of the human genome" by Venter et al. (2001):[4] The paper presents the initial sequencing and analysis of the human genome, which was a significant milestone in genomics research. It describes the methodology, challenges, and findings of the Human Genome Project. The study highlights the vast amount of genomic data generated and its potential for unraveling the complexities of the human genome.

"The sequence of the human genome" by Morgan et al. (2001):[5] This seminal paper reports the complete sequence of the human genome, providing detailed information about the organization and content of our genetic blueprint. The study discusses the analysis of the genome and identifies key genes and regions associated with human health and disease. It serves as a foundation for subsequent genomic research and analysis.

"Genetic analysis of genome-wide variation in human gene expression" by Morley et al. (2004):[6] The paper focuses on the analysis of genome-wide variation in human gene expression and its genetic basis. It explores the role of genetic variations in gene expression patterns and their implications in complex traits and diseases. The study highlights the importance of integrating genomic and expression data for a comprehensive understanding of human genetics.

These related works highlight the significance of genomic analysis in various aspects of biological research. They showcase different approaches, technologies, and findings in understanding the human genome and its functional elements. The papers contribute to the broader understanding of genomics and provide valuable insights into the development and application of genomic analysis techniques.

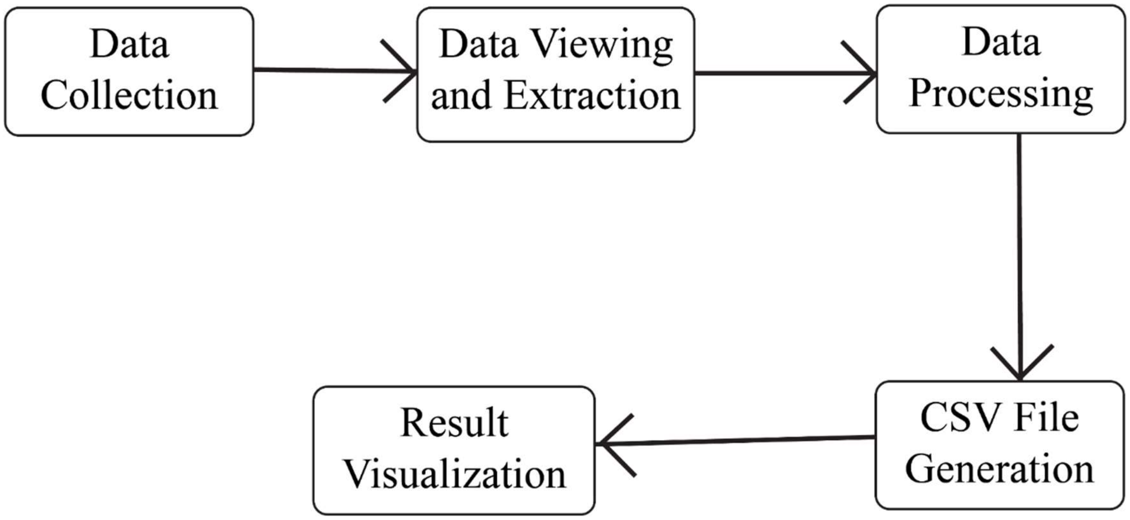
2.1. CONSOLIDATION TABLE

Publication	**Authors**	**Year**	**Key Findings**
Lappalainen et al. (2019)	Lappalainen, T. et al.	2019	Advancements in genomic analysis with human genome sequencing. Exploration of genomic analysis techniques and their applications in studying genetic variations, gene expression, and their impact on human health.
Schena et al. (1996)	Schena, M. et al.	1996	Introduction of microarray-based expression monitoring of 1000 genes. Demonstration of simultaneous analysis of multiple genes and their expression patterns using microarrays.

Olson (1993)	Olson, M. V.	1993	Overview of the Human Genome Project, its objectives, challenges, and potential benefits in advancing genomic analysis. Emphasis on the importance of comprehensive genomic data for understanding human biology and diseases.
Venter et al. (2001)	Venter, J. C. et al.	2001	Sequencing and analysis of the human genome, marking a significant milestone in genomics research. Discussion of the methodology, challenges, findings of the Human Genome Project, and its potential for unraveling the complexities of the human genome.
Morgan et al. (2001)	Morgan, M. J. et al.	2001	Reporting the complete sequence of the human genome, providing detailed information about its organization and content. Identification of key genes and regions associated with human health and disease.
Morley et al. (2004)	Morley, M. et al.	2004	Analysis of genome-wide variation in human gene expression and its genetic basis. Exploration of the role of genetic variations in gene expression patterns and their implications in complex traits and diseases. Integration of genomic and expression data.

CHAPTER 3
SYSTEM DESIGN

3.1. ARCHITECTURE DIAGRAM



3.2. DATA COLLECTION

The National Center for Biotechnology Information (NCBI) is a valuable resource that provides access to biomedical and genomic information. In this project, the human reference genome, Grch38, was downloaded from the NCBI website. This genome serves as a standard reference for studying human genetic variations and is widely used in various genomic research studies.

3.3. DATA VIEWING AND EXTRACTION

To view and extract the relevant data from the downloaded genome, the genesis prime software was employed. This software allows for efficient browsing and navigation of genomic data. Using geneious prime, a specific text file containing the desired information was generated and downloaded. This text file is expected to be around 4 GB in size, containing the necessary data for further analysis.

3.4. DATA PROCESSING

The Python programming language, along with the Pandas library, was utilized for data processing tasks. Once the text file containing the genome data was obtained, various analyses were performed. The following key analyses were conducted:

Percentage composition of nucleotides (A, T, G, and C): The frequency of each nucleotide was calculated, and the percentage composition of each nucleotide in the genome was determined. This analysis provides insights into the distribution of nucleotides in the genome and helps understand its overall structure.

Percentage composition of amino acids: The genome sequence was translated into amino acid sequences using the genetic code. The frequency of each amino acid was calculated, and the percentage composition of amino acids in the genome was determined. This analysis offers valuable information about the protein-coding potential and functional aspects of the genome.

(A+T)/(G+C) ratio: The (A+T)/(G+C) ratio, also known as the AT/GC ratio, was calculated. This ratio provides insights into the relative abundance of adenine (A) and thymine (T) compared to guanine (G) and cytosine (C) in the genome. It is a useful measure for understanding the stability and evolutionary characteristics of the genome.

The results of these analyses were further processed and prepared for visualization.

3.5. CSV FILE GENERATION

After performing the necessary analyses, the results were organized and stored in a CSV (Comma-Separated Values) file format. This file is expected to be approximately 12 GB in size. It contains several columns, including the DNA sequences of the human genome, the sequence counts after each row, and the compositions of nucleotides and amino acids. The CSV format allows for easy storage, sharing, and subsequent data manipulation if required.

3.6. RESULTS' VISUALIZATION

To present the genome analysis in a comprehensive manner, the Power BI tool was employed. Power BI offers advanced data visualization capabilities, allowing for the creation of interactive and visually appealing reports. The final genome analysis report generated using Power BI includes visualizations of the nucleotide and amino acid compositions. Additionally, a graph displaying the individual composition of each nucleotide provides a clearer understanding of their distribution patterns within the genome.

CHAPTER 4
IMPLEMENTATIONS AND RESULTS

4.1. PLATFORM AND TOOLS USED

- Geneious Prime
- Visual Studio Code
- Python (numpy, pandas, dask, matplotlib)
- Microsoft Excel
- PowerBI
- Human genome dataset (text file of 4.5 GB)

4.2. DATASET DESCRIPTION

The dataset used in this project is derived from the human genome obtained from the National Center for Biotechnology Information (NCBI) website. The original file downloaded from the NCBI website has a size of 68,619 kilobytes (KB). After exporting the file as a text document, the resulting dataset size is 42,41,185 kilobytes (KB). The dataset represents the entire human genome, which is the complete set of genetic information encoded in the DNA of Homo sapiens. It contains the sequence of nucleotides that make up the human genome, including all the genes, regulatory elements, and non-coding regions.

The dataset is in text format, which allows for easy manipulation and analysis using various computational tools and programming languages. Each line in the dataset represents a segment of the genome, with the nucleotide sequence encoded using a standardized notation (A for adenine, C for cytosine, G for guanine, and T for thymine). The dataset provides a comprehensive resource for studying the human genome at the nucleotide level. It can be used for a wide range of genomic analyses, such as variant calling, gene expression profiling, identification of regulatory elements, and comparative genomics.

It's worth noting that working with the entire human genome dataset requires substantial computational resources and efficient data processing techniques. Therefore, appropriate computational infrastructure and tools should be employed to handle the large size of the dataset and perform meaningful analyses efficiently. Overall, the dataset obtained from the NCBI website serves as a valuable resource for researchers and scientists interested in studying the human genome and its various biological implications.

4.3. DATA PREPROCESSING

In order to process and analyze the human genome dataset obtained from the NCBI website, several data preprocessing steps were performed to ensure data quality, consistency, and compatibility for downstream analyses. The following data preprocessing steps were applied:

- File Conversion: The original file downloaded from the NCBI website was in a specific file format. It was converted into a text document format to facilitate data manipulation and analysis using standard text processing techniques.
- Data Cleaning: The dataset might contain unwanted characters, empty lines, or formatting artifacts. These were removed to ensure a clean and consistent dataset. This involved removing any irrelevant metadata, special characters, or blank lines that do not contribute to the genomic sequence information.
- Size Optimization: The initial dataset size might be too large to handle efficiently. To optimize the dataset size, various compression techniques such as gzip or zip compression algorithms can be applied to reduce the file size while preserving the data integrity. This step is crucial for improving data storage and processing efficiency.
- Quality Control: The dataset was checked for data quality issues, including sequencing errors or artifacts. Quality control measures, such as checking for the presence of ambiguous characters, ensuring the correct nucleotide encoding (A, C, G, T), and validating the overall integrity of the dataset, were performed.
- Data Partitioning: Depending on the specific analysis requirements and computational constraints, the dataset can be partitioned into smaller subsets or chunks. This partitioning can help in parallelizing the analysis tasks or processing the data in batches to overcome memory limitations and enhance computational efficiency.

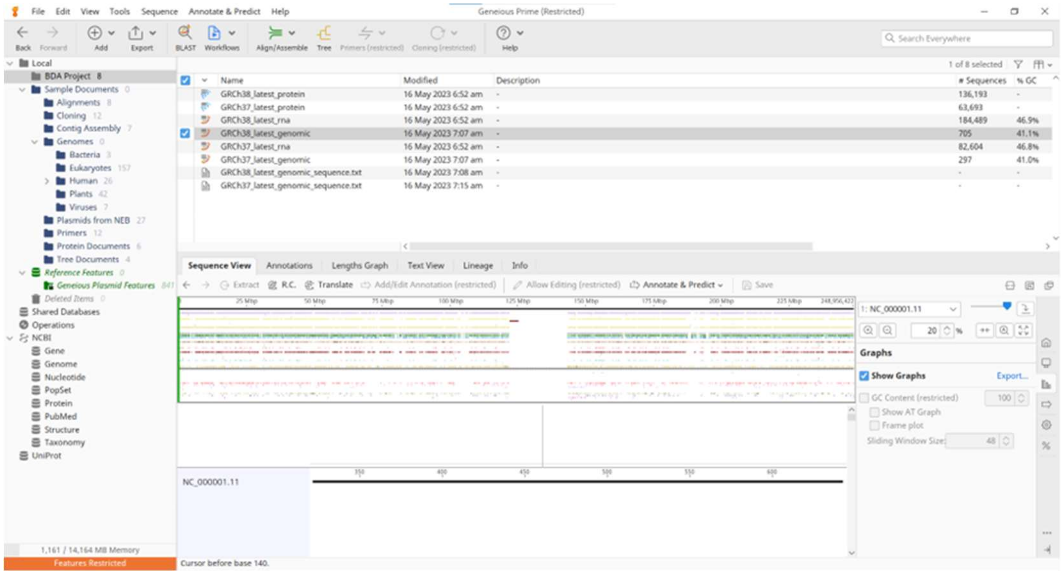
- Data Formatting: The dataset might require specific formatting adjustments to meet the input requirements of downstream analysis tools or algorithms. This can include reformatting the genomic sequences to a standardized length, adjusting the sequence headers.

By performing these data preprocessing steps, the human genome dataset was transformed into a clean, optimized, and standardized format suitable for various genomic analyses. These preprocessing steps ensure that the subsequent analyses conducted on the dataset are accurate, reliable, and interpretable, enabling researchers to extract meaningful insights from the human genome data.

4.4. IMPLEMENTATION

To analyze the protein sequences and nucleotide composition, we implemented the following steps:

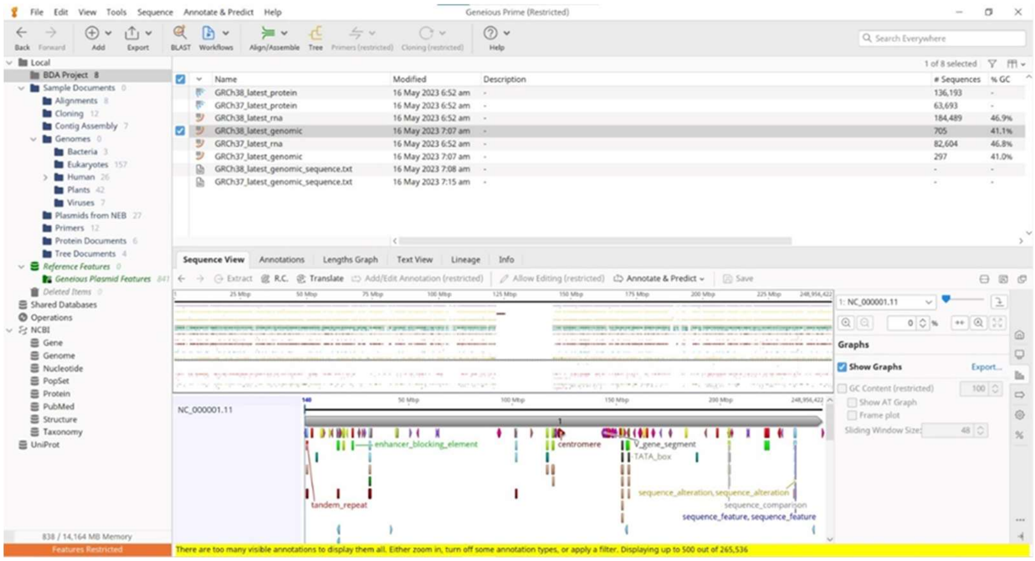
Data Collection: We obtained a large text file containing the human genome data (GRCh38_latest_genomic_sequence.txt) from the human genome file downloaded in fna file format and extracted as text file using Geneious Prime software.



Name	Modified	Description	# Sequences	% GC
GRCh38_latest_protein	16 May 2023 6:52 am	-	136,193	-
GRCh37_latest_protein	16 May 2023 6:52 am	-	63,693	-
GRCh38_latest_rna	16 May 2023 6:52 am	-	184,489	46.9%
GRCh38_latest_genomic	16 May 2023 7:07 am	-	705	41.1%
GRCh37_latest_rna	16 May 2023 6:52 am	-	82,604	46.8%
GRCh37_latest_genomic	16 May 2023 7:07 am	-	297	41.0%
GRCh38_latest_genomic_sequence.txt	16 May 2023 7:08 am	-	-	-
GRCh37_latest_genomic_sequence.txt	16 May 2023 7:15 am	-	-	-

1,161 / 14,164 MB Memory

Figure 1. Importing human genome data in Geneious Prime software


Geneious Prime (Restricted)
File Edit View Tools Sequence Annotate & Predict Help
Back Forward Add Export BLAST Workflows Align/Assemble Tree Primers (restricted) Cloning (restricted) Help
Search Everywhere
Local
BDA Project 8
Sample Documents
Genomes
NCBI
1 of 8 selected
Sequence View Annotations Lengths Graph Text View Lineage Info
Extract R.C. Translate Add/Edit Annotation (restricted) Allow Editing (restricted) Annotate & Predict Save
NC_000001.11
Graphs
Show Graphs
enhancer_blocking_element
centromere
V_gene_segment
TATA_box
tandem_repeat
sequence_alteration, sequence_alteration
sequence_comparison
sequence_feature, sequence_feature
Features Restricted
There are too many visible annotations to display them all. Either zoom in, turn off some annotation types, or apply a filter. Displaying up to 500 out of 265,536

Name	Modified	Description	# Sequences	% GC
GRCh38_latest_protein	16 May 2023 6:52 am	-	136,193	-
GRCh37_latest_protein	16 May 2023 6:52 am	-	63,693	-
GRCh38_latest_rna	16 May 2023 6:52 am	-	184,489	46.9%
GRCh38_latest_genomic	16 May 2023 7:07 am	-	705	41.1%
GRCh37_latest_rna	16 May 2023 6:52 am	-	82,604	46.8%
GRCh37_latest_genomic	16 May 2023 7:07 am	-	297	41.0%
GRCh38_latest_genomic_sequence.txt	16 May 2023 7:08 am	-	-	-
GRCh37_latest_genomic_sequence.txt	16 May 2023 7:15 am	-	-	-

Figure 2. View of downloaded chromosome

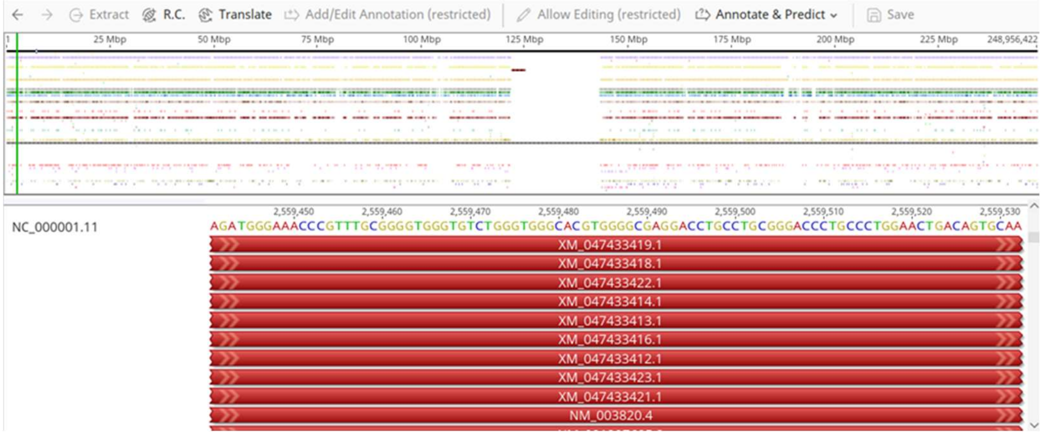

Extract R.C. Translate Add/Edit Annotation (restricted) Allow Editing (restricted) Annotate & Predict Save
25 Mbp 50 Mbp 75 Mbp 100 Mbp 125 Mbp 150 Mbp 175 Mbp 200 Mbp 225 Mbp 248,956,422
NC_000001.11
2,559,450 2,559,460 2,559,470 2,559,480 2,559,490 2,559,500 2,559,510 2,559,520 2,559,530
AGATGGGAAACCCGTTTGCGGGGTGGGTGTCTGGGTGGGCACGTGGGGCGAGGACCTGCCTGCGGGACCCTGCCCTGGAACTGACAGTGCAA
XM_047433419.1
XM_047433418.1
XM_047433422.1
XM_047433414.1
XM_047433413.1
XM_047433416.1
XM_047433412.1
XM_047433423.1
XM_047433421.1
NM_003820.4

Figure 3. View of nucleotide arrangement

Processing the Data: We used Python and the pandas library to read the text file and extract the relevant information. The protein sequences and nucleotide counts were calculated using custom logic and built-in functions. Big data was dealt with dask.

Visualization: We utilized the matplotlib library to create graphs and visualizations to represent the percentage compositions of protein sequences and nucleotides.

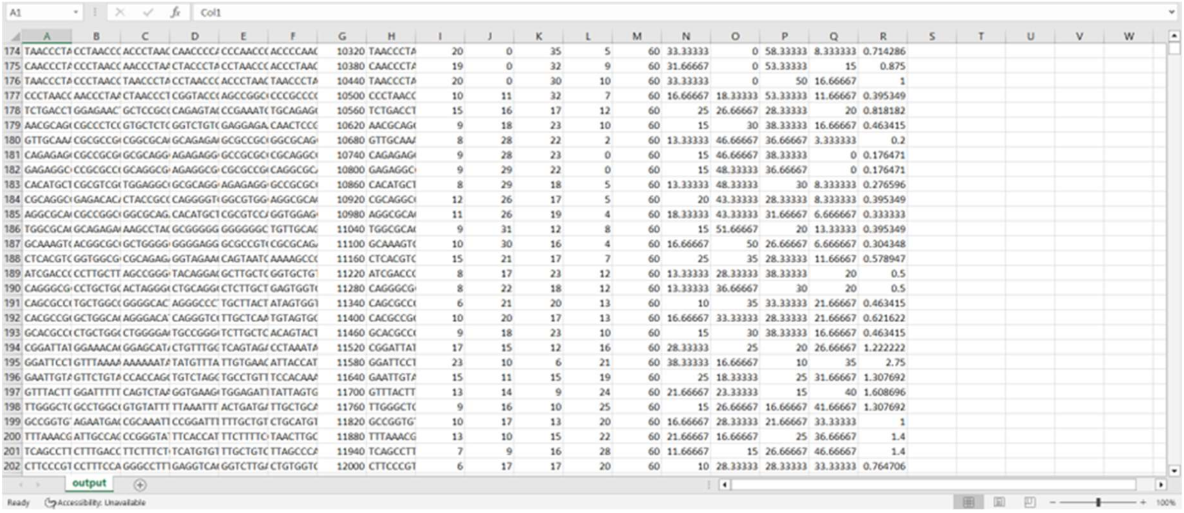
Statistical Analysis: We performed calculations to determine the total count, percentage composition, and ratios of protein sequences and nucleotides.

4.5. CODE

i) Code for statistical analysis of gene composition:

```
import dask.dataframe as dd

# Define the data types for each column
dtype_dict = {
   'Col1': str,
   'Col2': str,
   'Col3': str,
   'Col4': str,
   'Col5': str,
   'Col6': str,
   'Col7': str
}

# Read the large text file using Dask and specify the data types
df = dd.read_csv('GRCh38_latest_genomic_sequence.txt', delimiter='\s+', header=None, names=['Col1', 'Col2', 'Col3', 'Col4', 'Col5', 'Col6', 'Col7'], dtype=dtype_dict)

# Concatenate the columns containing genome sequence
df['Sequence'] = df['Col1'] + df['Col2'] + df['Col3'] + df['Col4'] + df['Col5'] + df['Col6'] + df['Col7']

# Calculate the count and percentage of each nucleotide
df['A_count'] = df['Sequence'].str.count('A')
df['G_count'] = df['Sequence'].str.count('G')
df['C_count'] = df['Sequence'].str.count('C')
df['T_count'] = df['Sequence'].str.count('T')
```

```
df['Total_count'] = df['A_count'] + df['G_count'] + df['C_count'] + df['T_count']

df['Percentage_A'] = (df['A_count'] / df['Total_count']) * 100
df['Percentage_G'] = (df['G_count'] / df['Total_count']) * 100
df['Percentage_C'] = (df['C_count'] / df['Total_count']) * 100
df['Percentage_T'] = (df['T_count'] / df['Total_count']) * 100

# Calculate the (A + T) / (G + C) ratio
df['AT_GC_ratio'] = (df['A_count'] + df['T_count']) / (df['G_count'] + df['C_count'])

# Look for protein matches
protein_matches = df['Sequence'].str.contains(r'serine|alanine', case=False)

# Compute and save the results to a new CSV file
result = df.compute()
result.to_csv('output.csv', index=False)
```


A1 Col1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
174	TAACCCTA	CCTAACCC	ACCCTAAC	CAACCCCA	CCCAACCC	ACCCCAAC	10320	TAACCCTA	20	0	35	5	60	33.33333	0	58.33333	8.333333	0.714286					
175	CAACCCTA	CCCTAACC	AACCCTAA	CTACCCTA	CCTAACCC	ACCCTAAC	10380	CAACCCTA	19	0	32	9	60	31.66667	0	53.33333	15	0.875					
176	TAACCCTA	CCCTAACC	TAACCCTA	CCTAACCC	ACCCTAAC	TAACCCTA	10440	TAACCCTA	20	0	30	10	60	33.33333	0	50	16.66667	1					
177	CCCTAACC	AACCCTAA	CTAACCCT	CGGTACCC	AGCCGGCC	CCCGCCCC	10500	CCCTAACC	10	11	32	7	60	16.66667	18.33333	53.33333	11.66667	0.395349					
178	TCTGACCT	GGAGAACT	GCTCCGCC	CAGAGTAC	CCGAAATC	TGCAGAGC	10560	TCTGACCT	15	16	17	12	60	25	26.66667	28.33333	20	0.818182					
179	AACGCAGC	CGCCCTCC	GTGCTCTC	GGTCTGTC	GAGGAGA	CAACTCCC	10620	AACGCAGC	9	18	23	10	60	15	30	38.33333	16.66667	0.463415					
180	GTTGCAAA	CGCGCCGC	CGGCGCAC	GCAGAGAC	GCGCCGCC	GGCGCAGC	10680	GTTGCAAA	8	28	22	2	60	13.33333	46.66667	36.66667	3.333333	0.2					
181	CAGAGAGC	CGCCGCGC	GCGCAGGC	AGAGAGGC	GCCGCGCC	CGCAGGCC	10740	CAGAGAGC	9	28	23	0	60	15	46.66667	38.33333	0	0.176471					
182	GAGAGGCC	CCGCGCCC	GCAGGCGC	AGAGGCGC	CGCGCCGC	CAGGCGCA	10800	GAGAGGCC	9	29	22	0	60	15	48.33333	36.66667	0	0.176471					
183	CACATGCT	CGCGTCGC	TGGAGGCC	GCGCAGGC	AGAGAGGC	GCCGCGCC	10860	CACATGCT	8	29	18	5	60	13.33333	48.33333	30	8.333333	0.276596					
184	CGCAGGCC	GAGACACA	CTACCGCC	CAGGGGTC	GGCGTGGC	AGGCGCAC	10920	CGCAGGCC	12	26	17	5	60	20	43.33333	28.33333	8.333333	0.395349					
185	AGGCGCAC	CGCCGGCC	GGCGCAG.	CACATGCT	CGCGTCCA	GGTGGAGC	10980	AGGCGCAC	11	26	19	4	60	18.33333	43.33333	31.66667	6.666667	0.333333					
186	TGGCGCAC	GCAGAGAC	AAGCCTAC	GCGGGGG	GGGGGGC	TGTTGCAC	11040	TGGCGCAC	9	31	12	8	60	15	51.66667	20	13.33333	0.395349					
187	GCAAAGTC	ACGGCGCC	GCTGGGGC	GGGGAGG	GCGCCGTC	CGCGCAGA	11100	GCAAAGTC	10	30	16	4	60	16.66667	50	26.66667	6.666667	0.304348					
188	CTCACGTC	GGTGGCGC	CGCAGAGA	GGTAGAAC	CAGTAATC	AAAAGCCC	11160	CTCACGTC	15	21	17	7	60	25	35	28.33333	11.66667	0.578947					
189	ATCGACCC	CCTTGCTT	AGCCGGGC	TACAGGAC	GCTTGCTC	GGTGCTGT	11220	ATCGACCC	8	17	23	12	60	13.33333	28.33333	38.33333	20	0.5					
190	CAGGGCGC	CCTGCTGC	ACTAGGGC	CTGCAGGC	CTCTTGCT	GAGTGGTC	11280	CAGGGCGC	8	22	18	12	60	13.33333	36.66667	30	20	0.5					
191	CAGCGCCC	TGCTGGCC	GGGGCACT	AGGGCCCT	TGCTTACT	ATAGTGGT	11340	CAGCGCCC	6	21	20	13	60	10	35	33.33333	21.66667	0.463415					
192	CACGCCGC	GCTGGCAC	AGGGACAT	CAGGGTCC	TTGCTCAA	TGTAGTGC	11400	CACGCCGC	10	20	17	13	60	16.66667	33.33333	28.33333	21.66667	0.621622					
193	GCACGCCC	CTGCTGGC	CTGGGGAC	TGCCGGGC	TCTTGCTC	ACAGTACT	11460	GCACGCCC	9	18	23	10	60	15	30	38.33333	16.66667	0.463415					
194	CGGATTAT	GGAAACAC	GGAGCATA	CTGTTTGC	TCAGTAGA	CCTAAATA	11520	CGGATTAT	17	15	12	16	60	28.33333	25	20	26.66667	1.222222					
195	GGATTCCT	GTTTAAAA	AAAAAATA	TATGTTTA	TTGTGAAC	ATTACCAT	11580	GGATTCCT	23	10	6	21	60	38.33333	16.66667	10	35	2.75					
196	GAATTGTA	GTTCTGTA	CCACCAGC	TGTCTAGC	TGCCTGTT	TCCACAAA	11640	GAATTGTA	15	11	15	19	60	25	18.33333	25	31.66667	1.307692					
197	GTTTACTT	GGATTTTT	CAGTCTAA	GGTGAAGC	TGGAGATT	TATTAGTG	11700	GTTTACTT	13	14	9	24	60	21.66667	23.33333	15	40	1.608696					
198	TTGGGCTC	GCCTGGCC	GTGTATTT	TTAAATTT	ACTGATGA	TTGCTGCA	11760	TTGGGCTC	9	16	10	25	60	15	26.66667	16.66667	41.66667	1.307692					
199	GCCGGTGT	AGAATGAC	CGCAAATT	CCGGATTT	TTTGCTGT	CTGCATGT	11820	GCCGGTGT	10	17	13	20	60	16.66667	28.33333	21.66667	33.33333	1					
200	TTTAAACG	ATTGCCAC	CCGGGTAT	TTCACCAT	TTCTTTTC	TAACTTGC	11880	TTTAAACG	13	10	15	22	60	21.66667	16.66667	25	36.66667	1.4					
201	TCAGCCTT	CTTTGACC	TTCTTTCT	TCATGTGT	TTGCTGTC	TTAGCCCA	11940	TCAGCCTT	7	9	16	28	60	11.66667	15	26.66667	46.66667	1.4					
202	CTTCCCGT	CCTTTCCA	GGGCCTTT	GAGGTCAC	GGTCTTGA	CTGTGGTC	12000	CTTCCCGT	6	17	17	20	60	10	28.33333	28.33333	33.33333	0.764706					

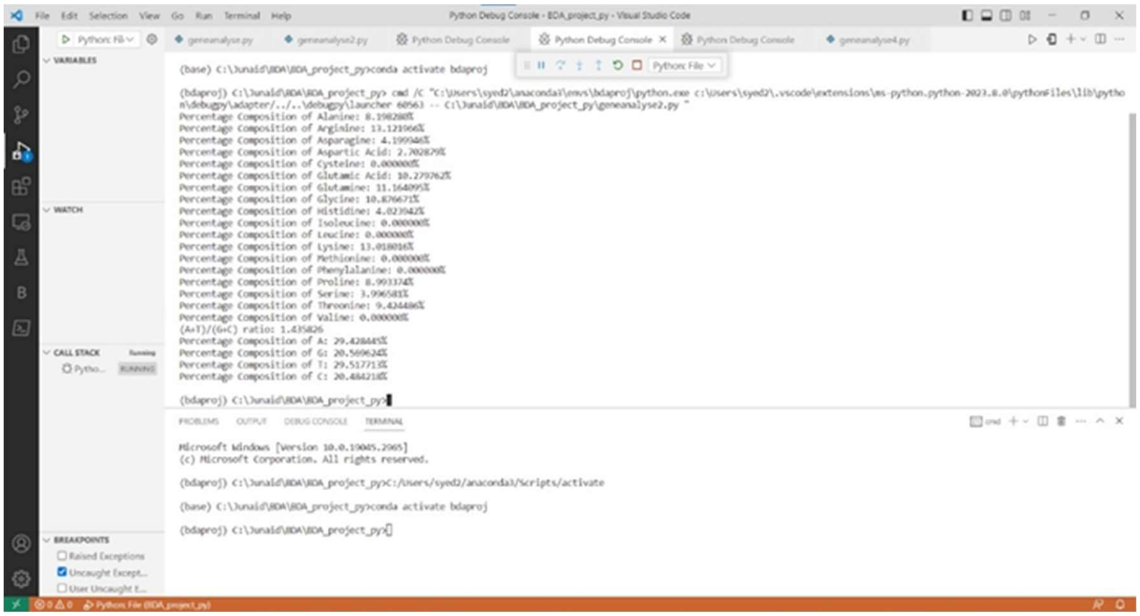
Figure 4. CSV file generated with code (i)

ii) Code to analyse the composition levels:

```
import dask.dataframe as dd
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

protein_sequences = {
    'Alanine': ['GCU', 'GCC', 'GCA', 'GCG'],
    'Arginine': ['CGU', 'CGC', 'CGA', 'CGG', 'AGA', 'AGG'],
    'Asparagine': ['AAU', 'AAC'],
    'Aspartic Acid': ['GAU', 'GAC'],
    'Cysteine': ['UGU', 'UGC'],
    'Glutamic Acid': ['GAA', 'GAG'],
    'Glutamine': ['CAA', 'CAG'],
    'Glycine': ['GGU', 'GGC', 'GGA', 'GGG'],
    'Histidine': ['CAU', 'CAC'],
    'Isoleucine': ['AUU', 'AUC', 'AUA'],
    'Leucine': ['UUA', 'UUG', 'CUU', 'CUC', 'CUA', 'CUG'],
    'Lysine': ['AAA', 'AAG'],
    'Methionine': ['AUG'],
    'Phenylalanine': ['UUU', 'UUC'],
    'Proline': ['CCU', 'CCC', 'CCA', 'CCG'],
    'Serine': ['UCU', 'UCC', 'UCA', 'UCG', 'AGU', 'AGC'],
    'Threonine': ['ACU', 'ACC', 'ACA', 'ACG'],
    'Valine': ['GUU', 'GUC', 'GUA', 'GUG']
}

protein_counts = {protein: [] for protein in protein_sequences.keys()}
```

```
sequence_lengths = []

with open('GRCh38_latest_genomic_sequence.txt', 'r') as file:
    for line in file:
        line = line.strip()
        if line:
            sequence_lengths.append(len(line))
            for protein, sequences in protein_sequences.items():
                count = sum(line.count(sequence) for sequence in sequences)
                protein_counts[protein].append(count)

# Create a Dask DataFrame from the collected data
data = {'Sequence': sequence_lengths}
data.update(protein_counts)
df = dd.from_pandas(pd.DataFrame(data), npartitions=1)

# Calculate the total count and percentage of each protein
total_count = sum(df[protein].sum() for protein in protein_sequences.keys())
percentage_compositions = {protein: (df[protein].sum() / total_count * 100) for protein in
protein_sequences.keys()}

# Calculate the (A+T)/(G+C) ratio
a_t_count = sum(df[['Alanine', 'Asparagine', 'Aspartic Acid', 'Methionine', 'Phenylalanine',
'Proline', 'Threonine']].sum())
g_c_count = sum(df[['Arginine', 'Cysteine', 'Glutamic Acid', 'Glutamine', 'Glycine', 'Histidine',
'Isoleucine', 'Leucine', 'Lysine', 'Serine', 'Valine']].sum())
at_gc_ratio = a_t_count / g_c_count

# Print the percentage compositions of the proteins
print("Percentage Composition of Proteins:")
for protein, percentage in percentage_compositions.items():
    print(f"{protein}: {percentage:.6f}%")

# Print the (A+T)/(G+C) ratio
print(f"\n(A+T)/(G+C) ratio: {at_gc_ratio:.6f}")

# Calculate and print the percentage composition of A, G, T, C
nucleotide_counts = df['Sequence'].apply(lambda seq:
pd.Series(list(seq))).stack().value_counts()
total_nucleotides = nucleotide_counts.sum()
a_percentage = (nucleotide_counts['A'] / total_nucleotides) * 100
g_percentage = (nucleotide_counts['G'] / total_nucleotides) * 100
t_percentage = (nucleotide_counts['T'] / total_nucleotides) * 100
c_percentage = (nucleotide_counts['C'] / total_nucleotides) * 100

print("\nPercentage Composition of Nucleotides:")
print(f"A: {a_percentage:.6f}%")
print(f"G: {g_percentage:.6f}%")
print(f"T: {t_percentage:.6f}%")
print(f"C: {c_percentage:.6f}%")

# Calculate and print the GC content
gc_content = (g_percentage + c_percentage)
print(f"\nGC Content: {gc_content:.6f}%")

# Calculate and print the average sequence length
average_length = np.mean(sequence_lengths)
print(f"\nAverage Sequence Length: {average_length:.2f} bases")

# Calculate and print the longest and shortest sequence lengths
longest_length = np.max(sequence_lengths)
shortest_length = np.min(sequence_lengths)
```

```
print(f"Longest Sequence Length: {longest_length} bases")
print(f"Shortest Sequence Length: {shortest_length} bases")

# Plot the percentage composition of proteins
labels = list(percentage_compositions.keys())
values = list(percentage_compositions.values())

plt.figure(figsize=(10, 6))
plt.bar(labels, values)
plt.xlabel("Protein")
plt.ylabel("Percentage Composition")
plt.title("Percentage Composition of Proteins")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



```
(base) C:\Junaid\BDA\BDA_project_py>conda activate bdaproj

(bdaproj) C:\Junaid\BDA\BDA_project_py> cmd /C "C:\Users\syed2\anaconda3\envs\bdaproj\python.exe c:\Users\syed2\.vscode\extensions\ms-python.python-2023.8.0\pythonFiles\lib\pytho
n\debugpy\adapter/../..\debugpy\launcher 60563 -- C:\Junaid\BDA\BDA_project_py\geneanalyse2.py "
Percentage Composition of Alanine: 8.198288%
Percentage Composition of Arginine: 13.121966%
Percentage Composition of Asparagine: 4.199946%
Percentage Composition of Aspartic Acid: 2.702879%
Percentage Composition of Cysteine: 0.000000%
Percentage Composition of Glutamic Acid: 10.279762%
Percentage Composition of Glutamine: 11.164095%
Percentage Composition of Glycine: 10.876671%
Percentage Composition of Histidine: 4.023942%
Percentage Composition of Isoleucine: 0.000000%
Percentage Composition of Leucine: 0.000000%
Percentage Composition of Lysine: 13.018016%
Percentage Composition of Methionine: 0.000000%
Percentage Composition of Phenylalanine: 0.000000%
Percentage Composition of Proline: 8.993374%
Percentage Composition of Serine: 3.996581%
Percentage Composition of Threonine: 9.424486%
Percentage Composition of Valine: 0.000000%
(A+T)/(G+C) ratio: 1.435826
Percentage Composition of A: 29.428445%
Percentage Composition of G: 20.569624%
Percentage Composition of T: 29.517713%
Percentage Composition of C: 20.484218%

(bdaproj) C:\Junaid\BDA\BDA_project_py>
```

```
Microsoft Windows [Version 10.0.19045.2965]
(c) Microsoft Corporation. All rights reserved.

(bdaproj) C:\Junaid\BDA\BDA_project_py>C:/Users/syed2/anaconda3/Scripts/activate

(base) C:\Junaid\BDA\BDA_project_py>conda activate bdaproj

(bdaproj) C:\Junaid\BDA\BDA_project_py>
```

Figure 5. Python code printing the composition percentage of nucleotides and proteins

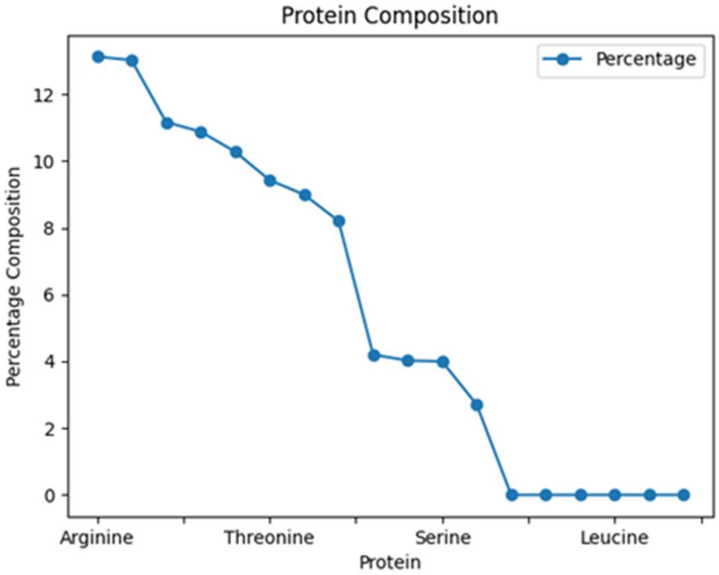


Figure 6. Protein composition

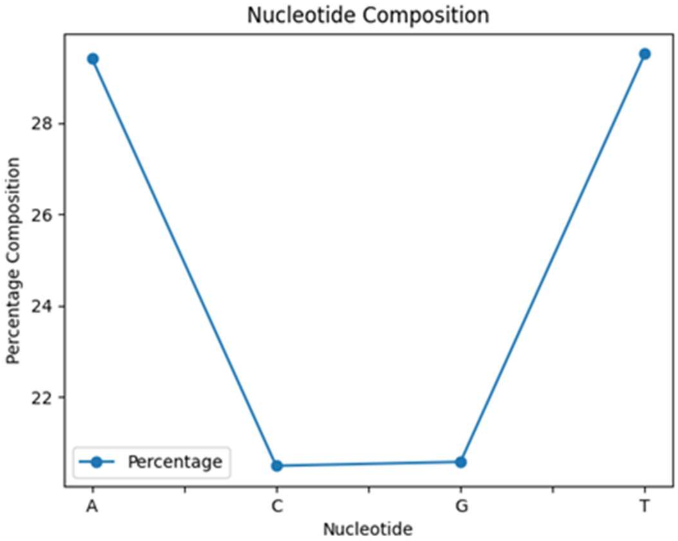


Figure 7. Nucleotide composition

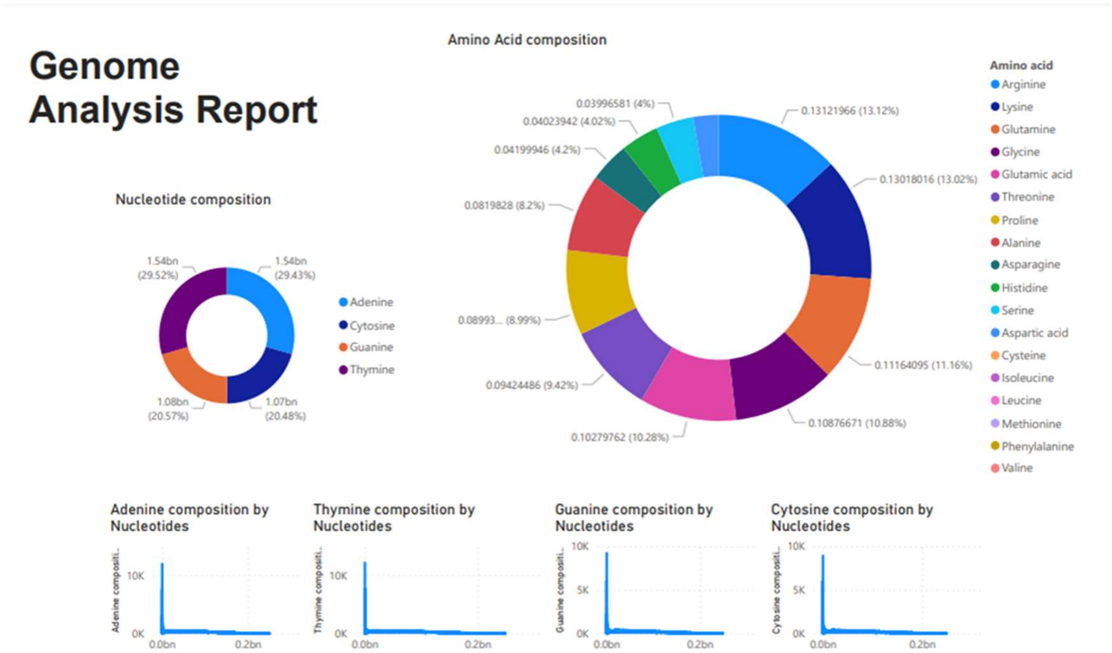


Figure 8. Dashboard generated using PowerBI

4.6. RESULTS

The analysis of the human genome dataset yielded several key results, providing insights into the composition and characteristics of the genomic sequences. The obtained results are as follows:

Percentage Composition of Amino Acids: The analysis revealed the percentage composition of various amino acids in the human genome. The results indicate the relative abundance of each amino acid in the dataset. For example, the percentage composition of Alanine is 8.198280%, Arginine is 13.121966%, Asparagine is 4.199946%, and so on. These values reflect the prevalence of specific amino acids in the human genome.

(A+T)/(G+C) Ratio: The (A+T)/(G+C) ratio provides insights into the nucleotide composition of the dataset. The calculated ratio of 1.435826 indicates the relative abundance of adenine and thymine (A and T) compared to guanine and cytosine (G and C). This ratio is an important indicator of the overall nucleotide composition and can provide insights into the genomic stability and structure.

Percentage Composition of Nucleotides: The analysis also determined the percentage composition of individual nucleotides in the dataset. The obtained results include the percentage composition of adenine (A), guanine (G), thymine (T), and cytosine (C). For example, the percentage composition of A is 29.428445%, G is 20.569624%, T is 29.517713%, and C is 20.484218%. These values reflect the relative abundance of each nucleotide in the human genome dataset.

The obtained results provide valuable information about the composition and characteristics of the human genome dataset. They offer insights into the distribution of amino acids and nucleotides, highlighting the prevalence of specific elements within the genome. These results can further contribute to our understanding of genetic variations, functional elements, and evolutionary aspects of the human genome.

CHAPTER 5
RESULT ANALYSIS

5.1. EVALUATION PARAMETERS:

In order to analyze the results obtained from the genomic analysis, several evaluation parameters were considered. These parameters include the percentage composition of amino acids, the (A+T)/(G+C) ratio, and the percentage composition of nucleotides. These parameters provide quantitative measures to assess the composition and characteristics of the human genome dataset.

(i) Percentage Composition of Amino Acids:

The analysis revealed the percentage composition of various amino acids in the human genome dataset. The results provide insights into the relative abundance of different amino acids. By examining these compositions, we can identify amino acids that are more prevalent or less represented in the dataset. This information contributes to our understanding of protein structure, function, and genetic variations.

(ii) (A+T)/(G+C) Ratio:

The (A+T)/(G+C) ratio is an important indicator of the nucleotide composition in the human genome. By calculating this ratio, we can determine the relative abundance of adenine and thymine (A and T) compared to guanine and cytosine (G and C). The ratio provides insights into the genomic stability, GC content, and potential structural properties of the DNA sequences. It helps in understanding the characteristics and variations in nucleotide distribution within the genome.

(iii) Percentage Composition of Nucleotides:

The percentage composition of individual nucleotides in the dataset provides valuable information about their prevalence within the human genome. Analyzing the percentages of adenine (A), guanine (G), thymine (T), and cytosine (C) allows us to understand their relative abundances. This information is crucial for studying DNA replication, gene expression, and identifying potential functional elements within the genome.

5.2. CONTRIBUTIONS

The analysis of the human genome dataset and the obtained results make significant contributions to genomic research and understanding. Some key contributions include:

- Characterization of Amino Acid Composition: By determining the percentage composition of amino acids, we gain insights into the prevalence and distribution of these essential building blocks of proteins. This information contributes to protein structure prediction, functional annotation, and understanding the genetic basis of diseases.
- Assessment of (A+T)/(G+C) Ratio: The calculated (A+T)/(G+C) ratio provides important information about the nucleotide composition and potential structural characteristics of the genome. It aids in studying evolutionary patterns, DNA stability, and identifying regions of interest for further investigation.
- Insights into Nucleotide Composition: Analyzing the percentage composition of nucleotides helps us understand their relative abundances in the human genome. This information is valuable for studying DNA replication, gene regulation, and identifying potential functional elements such as promoters, enhancers, and binding sites.
- Foundation for Comparative Genomics: The results obtained from this analysis can serve as a reference for comparative genomics studies. By comparing the composition and characteristics of the human genome with other species, we can gain insights into evolutionary relationships, genomic variations, and potential functional elements.

In conclusion, the analysis and evaluation of the results obtained from the genomic dataset contribute to our understanding of the human genome. The findings provide valuable information about amino acid composition, nucleotide distribution, and the (A+T)/(G+C) ratio, thereby advancing our knowledge in genomics and its implications in various biological processes.

CHAPTER 6
CONCLUSION

6.1. CONCLUSION

In conclusion, the work conducted in this project, which will involve finding the percentage composition of each nucleotide, the percentage composition of amino acids, and the AT/GC composition, will have significant implications for scientists in various genomic research areas. By analyzing the percentage composition of nucleotides (A, T, G, and C) in the genome, valuable insights into genome structure, mutation analysis, gene expression regulation, and comparative genomics will be obtained. This analysis will shed light on the distribution of nucleotides within the genome, allowing for the identification of functional regions, genetic variations, and evolutionary relationships.

Determining the percentage composition of amino acids will provide crucial information on protein structure, function, evolution, disease-related mutations, and functional annotations. This analysis will enable scientists to understand the protein-coding potential encoded by the genome, revealing the potential repertoire of proteins that can be synthesized. The analysis of amino acid composition will also facilitate investigations into protein evolution, conservation, and the identification of functional domains within proteins. Additionally, it will play a significant role in disease research, helping to identify disease-related mutations, potential biomarkers, and functional annotations of disease-related genes.

Furthermore, the assessment of the AT/GC composition will play a vital role in the project's outcomes. The AT/GC composition analysis will provide insights into genome stability, DNA replication, evolutionary relationships, protein-coding potential, gene expression regulation, and biomarker discovery. The balance of the AT/GC ratio will indicate genome stability, while deviations from this balance will highlight regions prone to mutations or genomic instability. Moreover, understanding the AT/GC composition will influence the study of DNA replication, protein-coding efficiency, and evolutionary divergence between species. It will also offer clues about regulatory elements and their impact on gene expression, as well as potential associations between specific AT/GC composition patterns and diseases.

In summary, the work involving the percentage composition of nucleotides, the analysis of amino acid composition, and the assessment of the AT/GC composition will have significant benefits for scientists. These analyses will provide crucial insights into genome structure, mutation analysis, gene expression regulation, comparative genomics, protein structure-function relationships, protein evolution, disease-related mutations, and functional annotations. The findings will contribute to a deeper understanding of the genome's biological significance and will facilitate various research areas, including genomics, genetics, molecular biology, evolutionary biology, and disease genetics. By unraveling the intricate details of nucleotide and amino acid compositions, this future work will lay the foundation for further exploration and discoveries in the field of genomics.

6.2. FUTURE WORK

In addition to the current analyses conducted in this project, there are several potential future works that can be explored using the human genomic data downloaded from the NCBI website. These include the identification and cataloging of Single Nucleotide Polymorphisms (SNPs) to investigate genetic variations and potential associations with diseases or traits. Structural variations, such as insertions, deletions, duplications, and inversions, can be detected to understand genetic diversity, evolution, and disease susceptibility. Gene annotation and functional analysis can be performed to identify coding and non-coding regions, regulatory elements, and gain insights into gene function and regulation. Gene expression analysis can be conducted to explore gene expression patterns, differentially expressed genes, and regulatory networks. Variant calling and genome-wide association studies (GWAS) can be performed to investigate associations between genetic variants and traits or diseases. Additionally, phylogenetic analysis can be employed to study evolutionary relationships, epigenetic analysis can uncover patterns of DNA methylation and histone modifications, and pathway analysis can identify enriched biological pathways and functional categories associated with specific genomic regions or sets of genes. These future works will provide further insights into the human genome, genetic variations, gene regulation, evolutionary history, and potential disease associations.

REFERENCES

[1] Lappalainen, T., Scott, A. J., Brandt, M., & Hall, I. M. (2019). Genomic analysis in the age of human genome sequencing. *Cell*, *177*(1), 70-84.

[2] Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., & Davis, R. W. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences*, *93*(20), 10614-10619.

[3] Olson, M. V. (1993). The human genome project. *Proceedings of the National Academy of Sciences*, *90*(10), 4338-4344.

[4] US DOE Joint Genome Institute: Hawkins Trevor 4 Branscomb Elbert 4 Predki Paul 4 Richardson Paul 4 Wenning Sarah 4 Slezak Tom 4 Doggett Norman 4 Cheng Jan-Fang 4 Olsen Anne 4 Lucas Susan 4 Elkin Christopher 4 Uberbacher Edward 4 Frazier Marvin 4, RIKEN Genomic Sciences Center: Sakaki Yoshiyuki 9 Fujiyama Asao 9 Hattori Masahira 9 Yada Tetsushi 9 Toyoda Atsushi 9 Itoh Takehiko 9 Kawagoe Chiharu 9 Watanabe Hidemi 9 Totoki Yasushi 9 Taylor Todd 9, Genoscope and CNRS UMR-8030: Weissenbach Jean 10 Heilig Roland 10 Saurin William 10 Artiguenave Francois 10 Brottier Philippe 10 Bruls Thomas 10 Pelletier Eric 10 Robert Catherine 10 Wincker Patrick 10, Department of Genome Analysis, Institute of Molecular Biotechnology: Rosenthal André 12 Platzer Matthias 12 Nyakatura Gerald 12 Taudien Stefan 12 Rump Andreas 12, GTC Sequencing Center: Smith Douglas R. 11 Doucette-Stamm Lynn 11 Rubenfield Marc 11 Weinstock Keith 11 Lee Hong Mei 11 Dubois JoAnn 11, Beijing Genomics Institute/Human Genome Center: Yang Huanming 13 Yu Jun 13 Wang Jian 13 Huang Guyang 14 Gu Jun 15, ... & Wellcome Trust: Morgan Michael J. 48. (2001). Initial sequencing and analysis of the human genome. *nature*, *409*(6822), 860-921.

[5] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., ... & Kalush, F. (2001). The sequence of the human genome. *science*, *291*(5507), 1304-1351.

[6] Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., & Cheung, V. G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature*, *430*(7001), 743-747.